Universidade de Aveiro
Ano 2010
Departamento de Electrónica Telecomunicações e Informática

Elcelina Rosa Correia Carvalho Silva

TÉCNICAS DE DATA E TEXT MINING PARA ANOTAÇÃO DE UM ARQUIVO DIGITAL

Universidade de Aveiro
Ano 2010
Departamento de Electrónica Telecomunicações e Informática

Elcelina Rosa Correia Carvalho Silva

TÉCNICAS DE DATA E TEXT MINING PARA ANOTAÇÃO DE UM ARQUIVO DIGITAL

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Electrónica e Telecomunicações – Especialização Sistemas de Informação, realizada sob a orientação científica do Doutor Joaquim Arnaldo Martins, Professor Catedrático e do Doutor José Manuel Matos Moreira, Professor Auxiliar, ambos do Departamento de Electrónica Telecomunicações e Informática da Universidade de Aveiro.

Com o apoio da Cooperação Portuguesa

Ao meu pai Fausto Silva, *In memoriam*

o júri

presidente

José Maria Amaral Fernandes
Professor Auxiliar da Universidade de Aveiro

José Manuel Matos Moreira
Professor Auxiliar da Universidade de Aveiro (Orientador)

Joaquim Arnaldo Carvalho Martins
Professor Catedrático da Universidade de Aveiro (Co-Orientador)

Fernando Joaquim Lopes Moreira
Professor Associado da Universidade de Portucalense

agradecimentos

Meus especiais agradecimentos aos meus orientadores, Doutor José Manuel Matos Moreira e Doutor Joaquim Arnaldo Martins pelas orientações e sugestões indispensáveis para a realização deste trabalho.

Agradeço também a todos os meus professores do Mestrado pelos conhecimentos que me transmitiram e aos colegas do curso.

Em terceiro lugar, agradeço à minha mãe, aos meus irmãos pela compreensão nos momentos que não estive presente.

Finalmente, mas não menos importante, deixo uma palavra de apreço ao Odair pelo companheirismo, ajuda e compreensão durante os momentos mais difíceis na feitura desta tese.

palavras-chave

Data Mining, Text Mining, Bases de Dados Textuais, Anotação de Documentos, Descritores de Texto.

resumo

O presente trabalho cujo Título é técnicas de *Data* e *Text Mining* para a anotação dum Arquivo Digital, tem como objectivo testar a viabilidade da utilização de técnicas de processamento automático de texto para a anotação das sessões dos debates parlamentares da Assembleia da República de Portugal.

Ao longo do trabalho abordaram-se conceitos como tecnologias de descoberta do conhecimento (KDD), o processo da descoberta do conhecimento em texto, a caracterização das várias etapas do processamento de texto e a descrição de algumas ferramentas *open souce* para a mineração de texto.

A metodologia utilizada baseou-se na experimentação de várias técnicas de processamento textual utilizando a *open source R/tm*. Apresentam-se, como resultados, a influência do pré-processamento, tamanho dos documentos e tamanhos dos corpora no resultado do processamento utilizando o algoritmo *knnflex.*

keywords

Data Mining, Text Mining, Textual Database, Document Annotation, Text Descriptor.

abstract

The present work whose title is “Techniques of Data and Text Mining for Annotation in a Digital Archive” has as its main objective to test the viability of using the techniques of automatic testing of texts for the annotation of the sessions of the debates in the National Assembly of the Republic of Portugal.

The work deals with concepts such as the techniques of discovering knowledge (KDD), the process of discovering knowledge in texts, the characterization of the various steps of processing a text and the description of some tools of open source for text mining.

The methodology used is an experiment of various techniques in text processing using the open source R/tm. The results show the influence of pre-processing, the size of the document and the size of the corpora in the results of the processing using the algorithm knnflex.

Conteúdo

Lista de Ilustrações

Lista de Gráficos

Lista de Tabelas

Capítulo 1

Introdução

A facilidade de publicação de informação na *Web* aliada à crescente produção de documentos de texto, faz com que actualmente, mais de 85% da informação disponível nas organizações esteja no formato de texto (Von, Mehler et al. 2005).

Muitas organizações que produzem ou que trabalham com um elevado número de documentos de texto, viram num primeiro momento as bibliotecas digitais como uma solução para disponibilizarem informações em forma de relatórios, artigos, teses, entre outros. Nos últimos anos, muitas dessas organizações começaram a sentir necessidade de técnicas mais sofisticadas para a estruturação dessas informações, tendo lançado a busca de técnicas automáticas para a categorização, o agrupamento e o resumo de documentos de modo a facilitar a pesquisa e extracção de informação a partir dos mesmos.

1.1 Motivação

A Assembleia da República Portuguesa, é uma instituição que tem vindo a apostar fortemente em meios tecnológicos para disponibilizar informações da actividade parlamentar. No âmbito deste investimento, foram desenvolvidos alguns projectos em parceria com a Universidade de Aveiro. Primeiramente foi desenvolvida a Biblioteca Digital, depois a Assembleia da República mostrou um interesse particular por um sistema de registo de vídeo das Sessões Parlamentares e Intervenções por cada Deputado em cada Sessão. Este sistema foi materializado com um outro sistema de Arquivo Audiovisual.

A Biblioteca Digital dos debates Parlamentares da Assembleia da República, segundo (Almeida, Martins et al. 2005), é um sistema que inclui a transcrição dos debates parlamentares e de todos os documentos relacionados com a actividade parlamentar, publicados desde a preparação da 1ª Constituição Portuguesa em 1821. Actualmente, a transcrição dos debates parlamentares para o diário da AR é disponibilizada na Internet cerca de 1 mês após cada Sessão Parlamentar.

O sistema de Arquivo Audiovisual da Assembleia da República (Almeida, Fernades et al. 2005), foi desenvolvido em 2006 com o objectivo de organizar, armazenar, indexar e permitir pesquisas de recortes de vídeo das sessões Parlamentares e Intervenções dos Deputados. Este sistema permite fazer o registo das intervenções com o preenchimento de atributos incluindo a intervenção do orador, o título, o assunto, os descritores (palavras-chave), a fase da sessão (antes da ordem do dia, ordem do dia ou votação), bem como o tempo correspondente ao início e ao final duma intervenção.

Normalmente, o processo da anotação dos vídeos decorre 2 ou 3 dias após a realização duma Sessão Parlamentar e é um processo moroso que envolve o preenchimento de um número elevado de atributos que, por sua vez, deve ser realizado por especialistas em catalogação e indexação de arquivos.

1.2 Objectivo

O objectivo deste trabalho consiste em testar a viabilidade da utilização de técnicas de processamento automático de texto para a anotação das sessões dos debates parlamentares da Assembleia da República. Em particular, pretende-se aplicar um conjunto de técnicas de treino e de análise de textos previamente anotados pelos especialistas em catalogação e indexação dos arquivos da AR, e testar a possibilidade de determinar automaticamente quais são os descritores (palavras-chave) que deverão ser associados a cada texto analisado.

1.3 Estrutura

A presente dissertação encontra-se estruturada em 7 Capítulos:

No primeiro e presente capítulo faz-se a contextualização da dissertação, fazendo referência à motivação, ao objectivo e à estrutura.

No segundo capítulo, **Tecnologias da Descoberta do Conhecimento**, contextualiza-se o processo da descoberta do conhecimento em bases de dados e a tecnologia de mineração em dados de diferentes formatos designadamente espacial, Web, multimédia incluindo vídeo, imagem, áudio e texto. No final do capítulo faz-se uma introdução à tecnologia *Text Mining*.

No terceiro capítulo, **Descoberta da Informação em Texto**, fala-se das bases de dados textuais, dos métodos indexação e recuperação de documentos e do processo da descoberta do conhecimento em texto.

No quarto capítulo, **Etapas de Processamento de texto**, explicam-se as três principais etapas de processamento do texto: Pré-processamento fazendo referência às técnicas de preparação de textos, Processamento mencionado os métodos de processamento de texto e Pós-Processamento indicando métodos de análise e avaliação dos resultados.

O quinto capítulo, **Ferramentas *Text Mining***, ilustram-se as várias ferramentas existentes para *text mining* e, através de alguns critérios de selecção, faz-se a escolha e caracterização de algumas numa perspectiva de escolher a mais adequada para utilizar no contexto do caso prático.

No sexto capítulo, **Estudo de Caso**, testa-se uma solução *text mining* e apresentam-se os resultados da classificação com o algoritmo *KnnFlex* em textos do Diário da Assembleia da República Portuguesa, submetidos às várias técnicas de pré-processamento. No final do capítulo faz-se uma avaliação das metodologias utilizadas e dos resultados conseguidos.

E, finalmente no sétimo capítulo, **Conclusões e Perspectivas**, faz-se as conclusões do Trabalho realizado e propõe-se futuras melhorias.

Capítulo 2

Tecnologias da Descoberta do Conhecimento

Cada vez mais aumenta a utilização de *Data Mining* tanto nas organizações públicas para detecção de fraudes, riscos, desperdícios ou ataques terroristas, como nas privadas particularmente bancos, companhias seguradoras, empresas da área da saúde, entre outras. A cada ano a quantidade da informação disponível duplica, sobretudo os documentos de texto, ao mesmo tempo que os custos de armazenamento estão a diminuir (Seifert 2004).

Neste capítulo pretende-se contextualizar a Tecnologia *Data Mining* e o processo da descoberta do conhecimento em bases de dados. O objectivo é mostrar que a tecnologia *Data Mining* pode ser aplicada a diferentes formatos de dados, designadamente espacial, *Web* e multimédia (texto, imagem, vídeo e áudio).

2.1 A Descoberta do Conhecimento em Bases de Dados

2.1.1 Tecnologia *Data Mining*

O processo da descoberta do conhecimento em bases de dados, começa com a determinação de metas e termina com o conhecimento descoberto. Como resultado, podem acontecer mudanças no domínio da aplicação, dando origem a novos repositórios de dados criados e ao início de um novo processo de descoberta do conhecimento (Maimon and Rokach 2005).

Data Mining é extracção ou "*mining*" de informação não trivial, não implícita, previamente desconhecida e potencialmente utilizável, em repositórios normalmente constituídos por uma grande quantidade de dados. Este processo é também chamado

descoberta do conhecimento em base de dados ou ***Knowlegde Discovery in Databases*** (**KDD**), (Zaiane 1999).

Autores como (Han and Kamber 2006) e (Maimon and Rokach 2005) defendem que *Data Mining* é apenas uma das etapas no vasto processo da descoberta do conhecimento em base de dados (**KDD**), incluindo os diversos tipos e fontes de dados (textual, multimédia, espacial, temporal, transaccional, *World Wide Web*), em que os processos em ordem progressiva incluem limpeza, integração, selecção, transformação de dados, descoberta de padrões de evolução e representação do conhecimento (Ilustração 1).

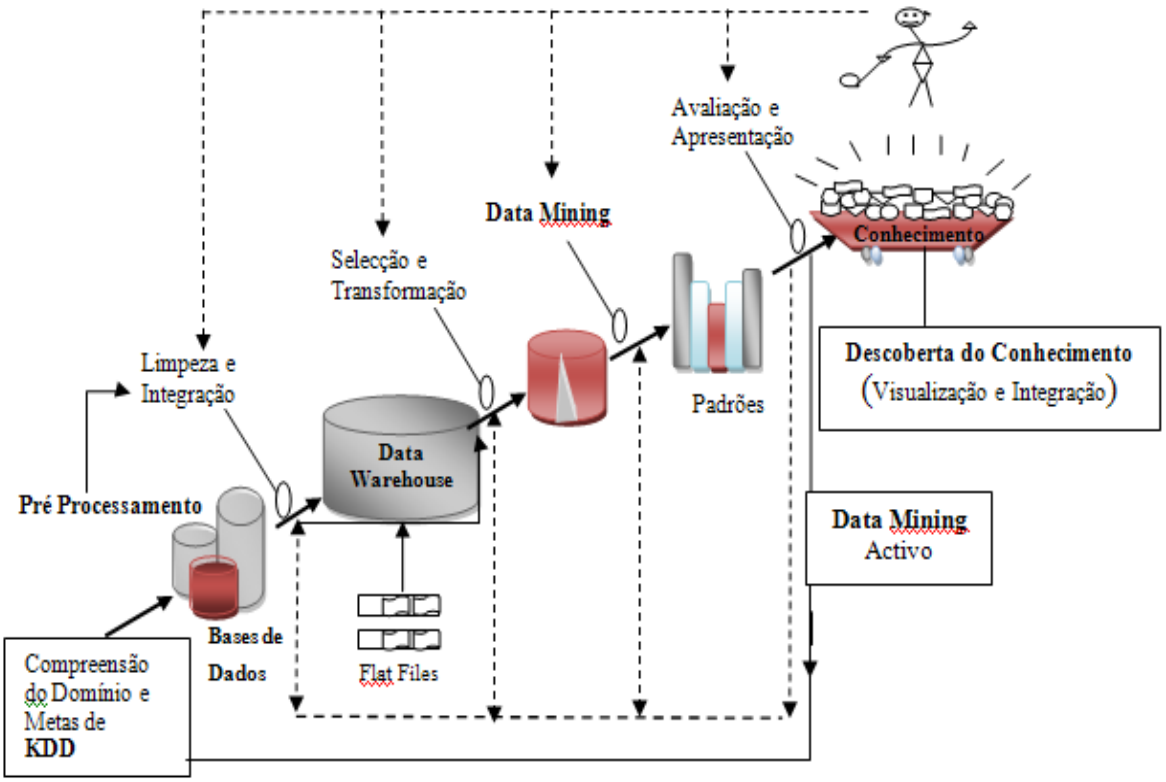


Ilustração 1 - Processo da descoberta do conhecimento em base de dados,

Fonte: Adaptação dos livros (Han and Kamber 2006) e (Maimon and Rokach 2005)

A **Compreensão do Domínio e Metas de KDD** é a etapa inicial do processo da descoberta do conhecimento em base de dados segundo (Han and Kamber 2006) porque, é nesta etapa que é definido o que se vai procurar e os objectivos a atingir ao longo do processo. Para isso, é necessário compreender o domínio (negócio), definir metas do utilizador final e o ambiente onde se vai processar a descoberta do conhecimento. Depois, começa-se o pré-processamento de dados.

O processo de **Limpeza e Integração** de dados, também designado por processo de Pré-Processamento, consiste na preparação dos dados relevantes que serão utilizados para a descoberta do conhecimento. A limpeza consiste na eliminação de dados duplicados ou irrelevantes para o processo, na eliminação de inconsistências nos dados e na integração de dados provenientes de diversas fontes, por exemplo, bases de dados, folhas de cálculo e documentos no formato XML. A integração envolve procedimentos como a conversão de formatos ou a uniformização dos dados (por exemplo, tratamento de datas em diferentes formatos). Em muitos casos, este processo termina com a integração dos dados numa *Data Warehouse*. O processo de limpeza e integração de dados é considerado fundamental uma vez que todo o processo de descoberta de conhecimento é realizado com base nos dados seleccionados.

O processo de **Selecção e Transformação** de dados, envolve a selecção de dados relevantes e a transformação dos mesmos para um formato apropriado para a realização do processo de *Data Mining*. No processo da selecção, faz-se a redução de dados recorrendo a métodos estatísticos como a análise de principais componentes aplicados de acordo com cada contexto específico. O processo de transformação dos dados envolve tarefas como o agrupamento de valores numéricos em classes, sem comprometer a sua integridade.

O processo de **Mineração** de Dados ou ***Data Mining***, é a etapa essencial onde são aplicados métodos ou algoritmos inteligentes para extrair padrões sobre os dados (*Data Patterns*). Este processo consiste em identificar padrões relevantes que representam o

conhecimento baseado em medidas. Nesta etapa, são escolhidos os algoritmos de Classificação, Regressão e Agrupamento (*Clustering*). A opção por um dos tipos de algoritmos depende das metas a atingir. Depois da escolha dos algoritmos, selecciona-se o método a ser usado que poderá ser, redes neuronais, árvores de decisão ou outros.

O processo da **Representação do Conhecimento**, representa a etapa final da descoberta do conhecimento. Consiste na visualização de informações e aplicação de técnicas de representação adequadas para apresentar o conhecimento extraído "*Mined knowledge*" ao utilizador. O sucesso desta etapa depende obviamente das etapas anteriores.

2.1.2 Tipos de Tecnologia *Data Mining*

Data Mining envolve a utilização de ferramentas sofisticadas de análise de dados para a descoberta de informações previamente desconhecidas, padrões válidos e relações em grandes colecções de dados (Seifert 2004).

É uma área interdisciplinar, porque relaciona com outras áreas como bases de dados, estatística, aprendizagem máquina ou inteligência artificial (Han and Kamber 2006). Não é específico apenas para um tipo de dados, pode ser aplicado a qualquer tipo de repositório de informação mas, os algoritmos e o processo de extracção de conhecimento diferem quando são aplicados a tipos de dados diferentes (cada tipo corresponde a uma tecnologia *mining* específica).

2.1.2.1 *Data Mining Relacional*

Relacional Data Mining ou *Multirelacional Data Mining* procura descobrir conhecimentos e padrões a partir das relações entre tabelas numa base de dados relacional (Maimon and Rokach 2005).

Cada tabela tem um conjunto de atributos representados em colunas e linhas. As colunas representam os atributos e as linhas representam registos de informação. Um registo,

corresponde a um objecto ou a uma relação entre objectos, é identificado por um conjunto de valores associados aos atributos normalmente identificados por uma chave única.

Geralmente quando se fala de *Data Mining* faz-se referência a ***Data Mining* Relacional**, porque o modelo clássico da descoberta do conhecimento em bases de dados (**KDD**) toma como referência as bases de dados relacionais e as *data warehouses.*

2.1.2.2 *Data Mining Espacial*

O aumento na utilização de sistemas de informação geográfica e das bases de dados espaciais que lhes estão associadas conduziu à necessidade da descoberta automática de conhecimento a partir de dados geográficos (Han and Kamber 2006). *Data Mining* Espacial, segundo este autor, é o processo da descoberta do conhecimento e padrões potencialmente utilizáveis, em grandes repositórios de dados espaciais. A complexidade de dados espaciais, torna este tipo de *Data Mining* mais complexo em comparação com o *Data Mining* relacional, porque abrange informações representadas por entidades complexas como pontos, linhas ou polígonos, que são consideravelmente mais difíceis de tratar do que as informações que podem ser simplesmente representadas recorrendo, por exemplo, a dados numéricos.

2.1.2.3 *Web Mining*

Internet é um grande repositório de dados abarcando uma enorme colecção de documentos em formatos variados. *S*egundo (Han and Kamber 2006), *Web Mining* consiste na utilização das técnicas de *Data Mining* para extrair, recuperar e analisar informação para a descoberta do conhecimento a partir de documentos e serviços da *web*.

Em *Web Mining* as informações podem ser retiradas a partir dos conteúdos das páginas Web, das estruturas das hiperligações e da análise da interacção de utilizadores com uma página Web (Magalhães 2008).

A extracção da informação a partir dos conteúdos das páginas *Web*, tem como objectivo extrair conteúdos em diferentes formatos, com o intuito integrar as informações e gerar conhecimentos.

Através das estruturas das hiperligações, pode-se categorizar as páginas Web com conteúdos semelhantes e gerar uma lista de páginas que podem ser úteis para comunidades virtuais com os mesmos interesses.

Com a análise da interacção de utilizadores com uma página *Web,* pode-se analisar os acessos dos utilizadores aos servidores através dos *logs* de acesso para extrair informações sobre a frequência de visita às páginas, comportamentos e interesses dos utilizadores. Estas Informações podem servir para melhorar os acessos, as personalizações das interfaces, os perfis dos utilizadores e a própria segurança do servidor.

2.1.2.4 Multimédia *Mining*

A crescente utilização da *Web* e a tecnologia de armazenamento de dados tem impulsionado a disponibilização de dados em diversos formatos. Pode-se dizer que Multimédia *Data Mining* é um processo de extracção automática de conhecimento a partir de documentos multimédia não estruturados como imagens, áudios, vídeos ou textos, recorrendo a métodos específicos para cada um desses formatos (Simoff, Djeraba et al. 2002).

1. Imagem

Imagens é uma das formas mais frequentes de comunicar e de transmitir a informação através da comunicação visual mas, muitas vezes exigem uma grande capacidade de armazenamento.

Consiste na extracção de padrões em grandes colecções de imagens com base em cor, conteúdo e, entre outros itens, através da aplicação de técnicas de *Data Mining* em imagem para a extracção de conhecimento implícito, isto é, não explicitamente armazenado na imagem (Han and Kamber 2006). Trabalha com valores representados em forma de *pixels* e consiste na aplicação de técnicas como reconhecimento de objectos, indexação de imagem, processamento de imagem, recuperação de imagem, bem como classificação, agrupamento, associação de imagens e reconhecimento de padrões.

2. Vídeo

A utilização de documentos audiovisuais vem crescendo nas últimas décadas, aumentando a cada dia o número de documentos disponíveis em formato audiovisual como *clips*, filmes, documentários, publicidade ou programas televisivos. Os canais televisivos também tem gerado um aumento considerável de documentos audiovisuais, porque actualmente produzem informações 24 horas por dia e milhares de horas por ano, ocupando terabytes de espaço em disco. O Vídeo *Mining* pode ser utilizado como uma solução para fazer pesquisa de vídeos sobre um determinado assunto ou acontecimento em estúdios de televisão.

3. Áudio

Áudio *Mining* interage com áreas de pesquisa como aprendizagem máquina, processamento de voz e algoritmos de processamento da linguagem. Extrai informações de grandes áudio *Warehouses* para a descoberta de conhecimento como informações, documentos, notícias ou diálogos numa determinada linguagem.

Segundo (Shetty and Achary 2008), existem dois tipos de Áudio *Mining*: Baseado em Indexação de texto e baseado em indexação de Fonema. O primeiro requer um processo de conversão onde cada palavra pronunciada em forma de voz é convertida em texto e em seguida estas palavras são identificadas num dicionário que pode conter milhares de palavras. O segundo trabalha apenas com som, analisando fonemas um a um, e procurando identificar o fonema correcto a partir dum dicionário de fonemas.

4. Texto

Nos últimos anos tem aumentado o número de documentos de texto nas organizações em que, actualmente existe muita informação relevante para a tomada de decisões em relatórios de texto. A mineração de texto (*Text Mining*) surge como uma solução para extrair informações importantes em documentos de texto.

Tratando-se de um tópico essencial para o desenvolvimento deste trabalho, este assunto será desenvolvido na secção seguinte.

2.1.3 A Tecnologia *Text Mining*

Num passado relativamente recente, as pesquisas nas áreas de *Data Mining* e da extracção de conhecimento foram focadas apenas nos tipos de dados estruturados, mas o aumento exponencial de documentos disponíveis nas organizações, despertou o interesse dos investigadores para a pesquisa de técnicas de manipulação e extracção de informação ou de conhecimento em dados não estruturados como os documentos de texto.

A descoberta do conhecimento em texto nasceu a partir da necessidade de descobrir informações, padrões e anomalias em textos de forma automática. Essa tecnologia, permite recuperar informações, extrair dados, resumir documentos, descobrir padrões, associar regras e realizar análises qualitativas ou quantitativas em documentos de texto (Aranha and Passos 2006).

Os textos podem ser estruturados, não estruturados ou semi-estruturados, incluindo emails, relatórios, resultados de pesquisas, páginas *Web*, campos textuais em base de dados ou documentos electrónicos e digitalizados.

Este processo envolve um grau de dificuldade significativo, considerando que as informações normalmente estão disponíveis em linguagem natural sem a preocupação com a padronização ou com a estruturação de dados (Monteiro, Gomes et al. 2006).

Para fazer a extracção de informação de bases de dados textuais "*Text Mining*", os métodos e as técnicas de *Data Mining* precisam ser integrados juntamente com técnicas

de recuperação de informações e construção ou utilização da hierarquia específica de dados de texto (dicionários e enciclopédias), sistemas de classificação de termos, palavras-chave ou conteúdos associados a descrição geral e concisa em documentos de texto, (Han and Kamber 2006) para descobrir informações relevantes entre pessoas, lugares, organizações de modo a classificar, agrupar, organizar e recuperar informações de documentos de acordo com os seus conteúdos (Aranha and Passos 2006).

Text Mining é vista como uma evolução de pesquisas em *Data Mining* aplicado ao texto. *Data Mining* e *Text Data Mining* têm semelhanças em termos de arquitecturas, dado que os dois sistemas têm rotinas de pré-processamento, algoritmos de descoberta de padrões, camada de representação dos resultados e ferramentas de visualização, (Feldman and Sanger 2007), utilização do algoritmo de agrupamento, associação e categorização (Baas 2008).

Data Mining aposta na previsão numérica enquanto, *Text Mining* aposta na extracção de informações relevantes, sumarização de texto, resposta a questões importantes e visualização de informação (Baas 2008). Uma outra diferença defendida por (Hearst 2003), é que em *Data Mining* extrai-se informação em bases de dados estruturadas desenhadas para que os programas façam processamento automático enquanto em *Text Mining* é texto em linguagem natural escrita para as pessoas lerem.

Text Mining é uma área interdisciplinar que não interage apenas com *Data Mining*, aprendizagem máquina e estatística mas também com linguagem computacional ou Processamento da linguagem natural (Waegel 2006).

A estatística é uma subárea da matemática que faz a análise de dados empíricos (Hotho, N¨urnberger et al. 2005). O contexto da estatística em relação a pesquisas de aplicações de *Text Mining* e sistemas de apoio à decisão (*Business Intelligence*), inclui técnicas de análise semântica em bio-informática, descoberta de acções jurídicas (*Jurisdictions*), plágio de publicações científicas e inquéritos de *helpdesk (Feinerer, Hornik et al. 2008).*

Estes autores afirmam que actualmente a maior parte das aplicações estatísticas trazem no pacote, funcionalidades de *Text Mining*.

O *Text Mining*, utiliza o processamento de linguagem natural, para estudar linguagem natural e a sua utilização pelo computador, especialmente técnicas de processar o texto de forma mais rápida que vai desde a manipulação de *strings* até o processamento da linguagem natural (Hotho, N¨urnberger et al. 2005).

2.2 Sumário

Este capítulo fez a apresentação dos principais conceitos relacionados com o processo da descoberta do conhecimento e a tecnologia *Data Mining*. Apresentou diferentes tipos de tecnologias *Mining,* os aspectos que têm em comum e as principais características que os distinguem. Abordou o conceito de *Text Mining* e as suas especificidades relativamente ao *Data Mining*.

Capítulo 3

Descoberta da Informação em Texto

O volume dos documentos tem aumentado consideravelmente nas organizações nos últimos anos, constituindo-se num importante repositório do conhecimento organizacional, mantendo registadas todas as informações pertinentes da organização. Neste contexto, o desafio está em extrair e correlacionar tais informações de modo a revelar conhecimentos para a tomada de decisões (Gonçalves, beppler et al. 2005).

Com este capítulo pretende-se contextualizar as bases de dados textuais e mecanismos de extracção e recuperação de informações nessas bases de dados. O objectivo é mostrar a necessidade da estruturação das bases de dados textuais para a extracção e recuperação de informações correctas em documentos de textos, uma vez que actualmente os documentos de texto aumentam a cada dia e as organizações precisam destas informações para tomar decisões.

3.1 Bases de dados Textuais

Repositórios de armazenamento de dados podem ser estruturados, não estruturados ou semi-estruturados. São estruturados quando têm uma estrutura lógica de representação como é o caso das bases de dados relacionais ou orientadas a objecto, *Data Werehouses*, *Flatfiles* ou folhas de cálculo. São semi-estruturados quando têm parcialmente uma estrutura de representação lógica, por exemplo, as páginas Web. E, não estruturados quando não têm nenhuma estrutura lógica de representação, como é o caso de muitos documentos de texto.

Um documento por sua vez pode também ser estruturado, não-estruturado ou semi-estruturado: É estruturado quando é especificado uma estrutura lógica, por exemplo documentos XML e artigos científicos. É não-estruturado quando não contém nenhuma estrutura lógica, por exemplo um texto. É semi-estruturado quando a sua estrutura lógica é parcialmente definida, por exemplo um *e-mail* (Khrouf and Soulé-Dupuy 2004).

Uma base de dados textual, é uma base de informação que contém uma colecção de documentos de texto pertencentes a vários domínios e várias categorias e, para extrair informações destas bases de texto, é necessário um esquema de classificação dos seus conteúdos (Yang 2005).

Um documento é uma unidade de dados textual discreto como relatórios de negócio, memorandos, *e-mail*, artigos científicos ou manuscritos, (Feldman and Sanger 2007). Pode ser também uma sequência de palavras, com pontuação, respeitando as regras gramaticais da linguagem, incluindo frases, parágrafos, secções, capítulos, livros, sítios *Web* e entre outros (Solka 2007). Contêm palavras descrevendo objectos que não são simples palavras-chave mas sim, frases longas ou parágrafos como especificação de produtos, relatórios de erros, mensagens de alertas, sumário de relatórios ou apontamentos (Han and Kamber 2006).

Desta forma, um documento existe dentro de um determinado contexto ou numa colecção particular (grupo de centenas, milhares ou até milhões de documentos). Pode ser membro de colecções ou de secções diferentes dentro de uma mesma colecção ou mesmo pertencer simultaneamente a várias colecções (Feldman and Sanger 2007). Este autor, defende que um documento tem as seguintes componentes:

- **Carácter:** Componente individual como letras, números, caracteres especiais, espaços, de entre outros, que são blocos construtores para níveis semânticos mais altos como palavras, termos e conceitos.
- **Palavra:** Conjunto de caracteres com um determinado significado, que só tem sentido quando utilizado dentro de um determinado contexto.

- **Termos:** São palavras simples ou frases compostas que normalmente caracterizam o assunto do documento.
- **Conceitos**: São características geradas por um documento pelo seu significado como palavras simples, compostas ou cláusulas que relacionam com identificação de termos específicos.

Muitas das bases de dados textuais estão em expansão e ainda não estão formalmente estruturadas porque a estrutura da informação de bases de dados textuais e a sintaxe por detrás dessas bases de dados variam de linguagem para linguagem (máquina e humana), de cultura para cultura, de utilizador para utilizador.

3.2 Indexação e Recuperação de Documentos

3.2.1 Indexação de Documentos

O processo de associar palavras-chave a documentos é designado por indexação (Feldman and Sanger 2007). Este processo procura descrever e caracterizar um documento com o auxílio de representações dos conceitos existentes no documento.

Segundo (Rose 2007), a indexação de documentos é o procedimento que mapeia um texto numa representação compacta, com palavras ou termos relevantes no texto do documento em conjunto com o seu peso (relevância). Estas palavras e frases-chave, são anexadas aos documentos para dar uma breve indicação dos conteúdos abordados nos mesmos (Witten 2005).

A indexação pode ser feita manualmente por profissionais de indexação (indexação manual) ou automaticamente através de algoritmos (indexação automática):
A indexação manual exige leitura do documento pelo profissional de indexação e identificação de termos importantes que caracterizam o documento. A indexação automática inclui extracção de frases mais significativas num documento, através de

frequência e associação de palavras recorrendo a um dicionário de palavras (Medelyan and Witten 2006).

Além da indexação manual e automática existe um outro tipo de indexação chamado de indexação de todo o texto (*full text indexing*) que indexa todas as palavras que ocorrem no texto (Pouliquen, Steinberger et al. 2003).

A indexação manual é uma tarefa subjectiva porque os profissionais normalmente fazem associação de termos de acordo com os seus próprios entendimentos em relação ao significado dos termos. Por conseguinte, palavras ou frases-chave indexadas manualmente por um único profissional de indexação poderão não conduzir a bons resultados devendo, nestes casos, ser feita por uma equipa de profissionais com uma visão comum e estandardizada dos significados dos termos (Witten, Medelyan et al. 2006).

A indexação automática por sua vez procura indexar documentos com termos que melhor descrevem o seu conteúdo através de métodos eficazes que despendem menos tempo.

É um processo que procura automaticamente termos que descrevem o conteúdo retratado num documento de forma a indexá-lo automaticamente. Este processo exige um conjunto de acções sobre o texto como a correcção ortográfica, a remoção de palavras pouco significativas (artigos, preposições, etc.) e, entre outras. Este conjunto de etapas é chamado de pré-processamento de texto. Após esta etapa são calculados os termos descritores para a indexação.

A indexação de documentos tem em vista facilitar o processo da recuperação de documentos de um arquivo (*corpus*) num sistema de busca ou recuperação de informação.

3.2.1 Recuperação de Documentos

A recuperação de Documentos é um processo que permite encontrar documentos que contêm um determinado conteúdo dentro de um grupo de documentos.

Dado um corpus de documentos e as necessidades de informação do utilizador especificadas usando um determinado conjunto de critérios de pesquisa, a recuperação de documentos é a tarefa que identifica documentos mais relevantes para as necessidades apontadas (Witten 2005).

Nas bibliotecas, livrarias ou arquivos tradicionais, este processo é realizado através de catálogos com a identificação do autor, título, classificação, palavras-chave, entre outros. A extracção automática dessas informações tem sido um dos propósitos de *Text Mining (Witten, Medelyan et al. 2006).*

A recuperação de documentos é muito utilizada nos motores de busca em que inserindo um texto como palavra-chave de busca, o sistema procura documentos que contêm essas palavras-chave e apresenta o resultado.

Existem sistemas que fazem a indexação de todas as palavras do documento "*bag of words*" e contabilizam a frequência de cada palavra no texto. Esta informação pode então ser usada na recuperação de informação, por exemplo, para resumir conteúdos ou extrair outras informações relevantes para os utilizadores. Desta forma, a recuperação de informação é muitas vezes vista como uma extensão da recuperação de documentos (Witten 2005).

Outro problema relacionado com a recuperação de documentos envolve o acesso a documentos com conteúdos semelhantes, por exemplo, documentos que retratam um assunto específico duma categoria pré-definida (desporto, política, cultura, entre outros). Uma das formas de resolver esse problema é a classificação dos conteúdos do documento em categorias pré-definidas e o agrupamento (*clustering*) de documentos com conteúdos semelhantes.

A categorização dos documentos é conhecida como uma forma de aprendizagem supervisionada para recuperação de documentos usando categorias pré-definidas e previamente conhecidas. Por seu lado, agrupamento de documentos é conhecido como aprendizagem não supervisionada porque não existe nenhuma categoria ou classe pré-definida mas, a recuperação de documentos faz-se através da criação de conjuntos de documentos semelhantes (Witten 2005).

Num sistema de *Text Mining,* a recuperação de documentos é normalmente interligada com a indexação, porque a própria recuperação faz-se através de procura de termos indexados.

3.3 O Processo da Descoberta do Conhecimento em Texto

A explosão da informação demanda que as organizações procurem novas capacidades no que concerne à análise de dados de modo a criar a informação necessária para a tomada da melhor decisão e obtenção de vantagens competitivas (Penteado and Boutin 2008).

Uma *Data Warehouse* é um repositório de dados recolhidos a partir de diversas fontes usado como uma base de dados unificada e integrada. Para isso, os dados passam por um conjunto de transformações em que primeiramente são carregados, limpados, transformados e integrados, dando origem a uma única base de dados "*data warehouse*" (Seifert 2004).

Uma *Textual Warehouse* ou *warehouse* de documentos é uma base de informação unificada a partir de várias bases de dados textuais orientados ao assunto (dados relevantes para a análise de informação). Essas bases de dados são filtradas, integradas, guardadas e organizadas para a extracção, recuperação, interrogação e análise multidimensional de dados (Khrouf and Soulé-Dupuy 2004).

Um sistema de *Text Mining* pela natureza dos resultados que apresenta (padrões, conexões e tendências) tem que ser dinâmico, pois deve ter uma interface que recebe uma colecção de documentos através do ciclo iterativo de entrada e saída com o utilizador Final. Este selecciona as suas opções através da inserção de critérios, os dados são processados e apresentados em vários tipos de saída como padrões, mapas ou tendências (Feldman and Sanger 2007).

Este autor defende que num nível mais abstracto da arquitectura dum sistema de descoberta do conhecimento em texto, o sistema de *Text Mining* está subdividido em 4 áreas: 1ª - Tarefa do pré-processamento; 2ª - Operações de Mineração "; 3ª - Camada de Apresentação, Componentes e Funcionalidade do Browser e 4ª - Refinamento dos Processos. Este sistema é apresentado na Ilustração 2.

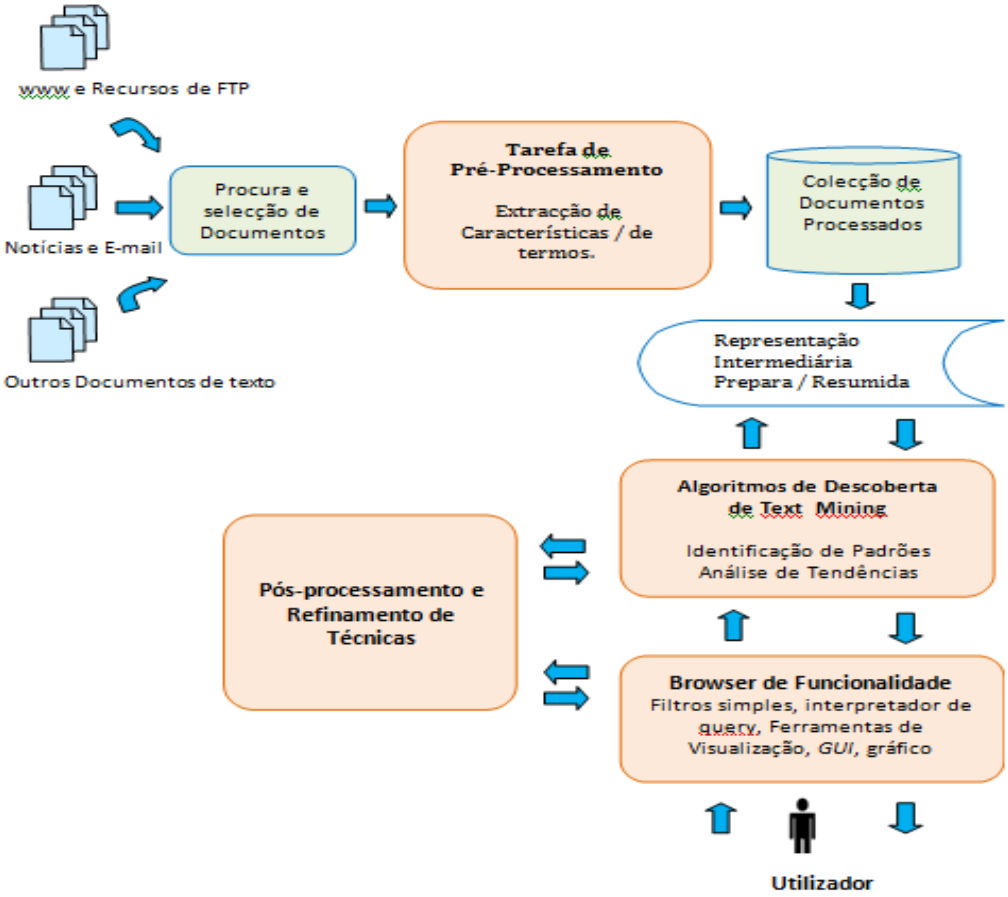


Ilustração 2 - Arquitectura para um sistema de descoberta do conhecimento

Adaptação do livro (Feldman and Sanger 2007).

A Tarefa do pré-processamento, inclui todas as rotinas, processos e métodos necessários para a preparação de dados a utilizar nas operações de descoberta do conhecimento. O pré-processamento converte os dados originais num formato adequado para aplicar os vários métodos de extracção do conhecimento. Esta etapa é considerada uma etapa crítica em qualquer sistema de *Text Mining*.

As Operações de Mineração são consideradas a parte principal dum sistema de descoberta de informação em texto (*Text Mining)*, dado que abarcam a descoberta de padrões e a análise de tendências através do cálculo da frequência de termos em documentos e da aplicação de algoritmos de descoberta do conhecimento em texto.

A Camada de Apresentação, Componentes e Funcionalidade do Browser, centra-se na interface gráfica com o utilizador incluindo uma linguagem para a interpretação dos pedidos dos utilizadores e a apresentação dos padrões extraídos através das funcionalidades do navegador (*browser)*.

O Refinamento dos Processos, tem como objectivo melhorar os resultados alcançados através da aplicação de novos métodos de preparação de dados, de aplicação de algoritmos e da análise de resultados.

3.4 Sumário

Neste capítulo, falou-se das bases de dados textuais, da indexação e recuperação de documentos de texto, da necessidade de criação de técnicas automáticas para indexar e anotar documentos de texto nas bases de dados textuais. Abordou-se o processo da descoberta de conhecimento em texto que será detalhado no próximo capítulo.

Capítulo 4

Etapas de Processamento de Texto

A pesquisa de informação vem sendo uma actividade comum para todas as pessoas que utilizam o computador e que acedem aos recursos da Internet no seu dia-a-dia mas, a formulação de uma pesquisa de termos adequada é o problema chave na pesquisa electrónica (Medelyne 2005). No caso dos repositórios de documentos no formato de texto um das formas de resolver este problema seria organizar os documentos em grupos, de acordo com as frases-chave e palavras-chave em que, estas corresponderiam aos principais conceitos presentes no conteúdo do documento.

Neste capítulo, almeja-se descrever as etapas do processamento de texto: Pré-processamento, Processamento e Pós-processamento. Visa essencialmente mostrar a diferença entre as várias etapas do processamento textual, deixando assim uma visão clara da importância de cada uma delas num sistema *Text Mining*.

4.1 Pré-Processamento de Texto

Para fazer a mineração ou extracção da informação em grandes colecções de textos é necessário o pré-processamento de documentos de texto e o armazenamento dessas informações sob a forma de dados estruturados, mais fáceis de processar do que um ficheiro de texto (Hotho, N¨urnberger et al. 2005).

A etapa do Pré-processamento de texto também chamada etapa de preparação do texto, visa essencialmente fazer a remoção dos dados desnecessários para o entendimento do texto e extracção do conhecimento (Monteiro, Gomes et al. 2006).

O processo de preparação do texto inclui um conjunto de acções sobre um texto, nomeadamente: correcção ortográfica, remoção de *stopwords*, lematização, definição de *n-grams,* cálculo de peso e selecção de palavras-chave.

4.1.1 Correcção Ortográfica

A correcção ortográfica dum documento visa eliminar possíveis erros ortográficos num texto através de um corrector ortográfico utilizado juntamente com um dicionário de línguas (nem todos os textos estão escritos na mesma língua), onde é comparada cada palavra do texto com os termos do dicionário. Caso houver uma coincidência entre as palavras, a mesma é considerada sintacticamente correcta, (Monteiro, Gomes et al. 2006). Por conseguinte, pode-se dizer que a correcção ortográfica tem como principal objectivo corrigir num texto palavras candidatas a serem termos a indexar.

Este processo contempla também a eliminação de *hifens*, pontuação, acentos, transformação de letras do texto em maiúsculas ou minúsculas, podendo deste modo, levar a perda de informação.

4.1.2 Remoção de *Stopwords*

Para reduzir o tamanho do dicionário de palavras e a dimensão dos documentos dentro duma colecção de documentos, reduz-se a lista de palavras que descrevem os documentos através de métodos como filtragem, lematização ou *stemming* (Hotho, N¨urnberger et al. 2005).

O método de filtragem é também conhecido como método de remoção de ***stopword*** ou remoção de ***stoplist*** que é a remoção de palavras como artigos, conjunções, preposições, entre outras. Estas palavras são comuns, repetem-se muitas vezes num texto e não acrescentam nem retiram informações relevantes, por exemplo: “O”, “a”, “e”, “mas”, ou, “para”, entre muitas outras (Solka 2007). Este processo também permite reduzir o

tamanho do documento indexado, deixando apenas palavras essenciais para a extracção da informação (Rose 2007).

4.1.3 Lematização

A Lematização ou *Stemming* é um outro método utilizado para diminuir a lista de palavras indexadas num documento (Waegel 2006).

Consiste na remoção de variações de palavras do tipo (plural, gerúndio, prefixos, sufixos, género e número) de modo que a palavra fique só com a *stem* (raiz) (Monteiro, Gomes et al. 2006).

Este método permite fazer com que palavras semelhantes fiquem com a mesma raiz, por exemplo, desenvolvido e desenvolvimento ficam desenvolv, o que, segundo (Waegel 2006) e (Rose 2007) tem dois efeitos positivos: primeiro, o número de termos indexados é reduzido porque termos similares são mapeados como uma única entrada; segundo, a relevância dos resultados é muitas vezes melhorada significativamente.

4.1.4 N-Grams

Quando se analisa uma palavra isoladamente, perde-se algum contexto da informação e, a solução desse problema passa pela aplicação do *n-grams,(Rose 2007).* Segundo este autor *n-grams* é a sequência duma palavra num comprimento *n.*

Na perespectiva de (Elberrichi and Aljohar 2007), N-grams é uma sequência de caracteres consecutivos existindo em **n** tamanhos em que **n** pode ser (n=1,2, … n). Cada tamanho pode ser designado por um nome: *Bi-grams* para 2 caracteres, *tri-grams* para 3 caracteres, *quadri-grams* e assim sucessivamente.

N-grams faz parte da identificação da linguagem porque as palavras são consideradas isoladas e é utilizada para criar uma sequência de palavras. No pré-processamento de

texto é utilizado para reorganizar um texto lematizado e sem *stopword* de modo a originar um novo texto com alguma sequência lógica compreensível, (Witten 2005).

4.2 Processamento de Texto

4.2.1 Cálculo de Frequência de Palavras (*weigth*)

O cálculo do peso dos termos num documento ou colecção de documentos é um factor importante na performance de um sistema de recuperação da informação, (Samat, Murad et al. 2008). Tem como objectivo reduzir a importância relativa do tamanho dos documentos relativamente ao número de ocorrência dos termos, (Waegel 2006).

Sabendo que uma colecção de ***d*** documentos é composta por ***t*** termos representado como uma matriz ***A*** de ***t*** **x** ***d*** (matriz de termos que aparecem num documento) em que, cada elemento da matriz representa um peso (*weigth*) de termos t_i em cada documento d_j.

Existem várias formas de calcular o peso ou a frequência de termos num documento. Uma das formas é a utilização da seguinte fórmula:

$$W_{ij} = L_{ij} G_i N_j$$

Em que, o peso $\mathbf{W_{ij}}$ do termo t_i no documento d_j é um produto de três factores onde:

- L_{ij}, representa o peso local do termo ***i*** no documento ***j***. Calcula o número de vezes que um termo aparece dentro dum documento. Se aparecer muitas vezes significa que este termo é muito pertinente para o documento.
- G_i, representa o peso global do termo ***i*** dentro da colecção de documentos (corpus). Calcula quantas vezes um termo aparece dentro de toda a colecção de documentos.
- N_j, é um factor de normalização para documento ***j***. É usado para corrigir discrepâncias em documentos com tamanhos diferentes.

Uma outra fórmula utilizada para calcular a frequência de termos é a seguinte:

$$tf - idf$$

A frequência do termo (*Term frenquency* - ***tf***), descreve a ocorrência dum termo num documento e o inverso da frequência do documento (*inverse document frequence* - ***idf***) conta o número de ocorrências do termo noutros documentos. Tanto ***tf*** como ***idf*** descrevem termos num documento e quais ocorrem maior número de vezes (Rose 2007).

4.2.2 Associação e Extracção de Termos e Frases-chave

Para localizar informações num determinado documento, é necessário um ambiente confortável para encontrar e ler documentos. Para isso, são necessários algoritmos que extraiam informações dos documentos sob a forma de palavras-chave com uma consistência semelhante às palavras-chave indexadas pelos profissionais que fazem a indexação manual (Medelyan and Witten 2006).

As palavras-chave são definidas como termos ou palavras que resumem e descrevem o conteúdo de um documento, enquanto frases-chave são multi-palavras (ou várias palavras) caracterizadas como frases resumo de documentos, usados para organizar bibliotecas e promover acesso a temáticas específicas (Medelyne 2005) e (Medelyan and Witten 2006).

A escolha de frases-chave e palavras-chave pode ser feita por profissionais de indexação de forma manual mas, o aumento crescente de documentos quer na Internet quer nas organizações tem se tornado insustentável a indexação manual de modo que indexação automática tem despertado muito interesse aos investigadores. É importante salientar que a associação manual de palavras-chave com elevada qualidade impõe elevados custos além de consumir muito tempo.

Existem dois métodos de indexação para a extracção automática de palavras ou frases-chave: Extracção de Palavras-chaves e Associação de Palavras-chave, (Pouliquen, Steinberger et al. 2003) e (Witten 2005).
A Extracção de Palavras-chave (*Keyword Extraction*) identifica palavras-chave presentes no texto que melhor caracterizam o conteúdo do documento. A extracção de frases-chave (*Keyphrases Extraction*), identifica frases que melhor caracterizam o assunto tratado num documento.

Existem também o conceito de associação de palavras-chave (*Keyword Assigment*), que identifica palavras-chave apropriadas ao conteúdo dum texto através dum dicionário de termos de referência (*thesaurus*) (Pouliquen, Steinberger et al. 2003). Neste método, os termos do dicionário são utilizados como descritores e não precisam estar necessariamente no texto. Este método oferece vantagens porque os termos já estão num dicionário de termos mas, também tem a desvantagem de não aceitar novos termos porque estes não se encontram no dicionário.

A qualidade da indexação automática mede-se através de palavras-chave associadas a documentos (termos indexados) que apresentam resultados aproximados dos da indexação manual e dos dum dicionário de termos de referência (*thesaurus*); caso contrário, os termos indexados podem não corresponder, na totalidade, o conteúdo abordado num documento.

Recentes investigações têm apontado que nem sempre palavras que ocorrem mais vezes num documento são as que representam melhor o seu conteúdo (Medelyan and Witten 2006). Para comprovar essa abordagem mede-se a similaridade e a relação dos termos abordados num documento e num conjunto de documentos.

4.2.3 Extracção de termos Similares

Muitas investigações já foram levadas a cabo para desenvolver métodos para descobrir termos similares (sinónimos) em corpus, páginas Web e dicionários de múltiplas línguas. Uma regra básica assumida em muitas destas pesquisas é que palavras similares são usadas num mesmo contexto, mas podem diferir de acordo com o lugar onde o contexto é definido (documentos ou contexto gramatical elaborado, entre outros) ou do lugar onde a função de similaridade é processada (Senellart and D. Blondel 2008). Este autor fala de alguns métodos de descoberta de termos similares:

4.2.3.1 Descoberta de palavras similares num Corpus extenso

Umas dos princípios básicos para a descoberta de palavras similares num corpus extenso de documentos de texto é a assunção de que palavras são similares são usadas num mesmo contexto (Senellart and Vincent 2008).

Existem muitas técnicas de extracção de termos similares através da análise lexical: Uma das técnicas utilizadas é a aplicação do contexto gramatical que pode ser visto como um tripleto (**w, r, w'**) em que **w** e **w'** são duas palavras e **r** caracteriza a relação entre essas duas palavras. Essa relação pode ser do tipo: Um adjectivo modifica um nome, um nome modifica um nome, um nome é sujeito do verbo, um nome é objecto directo dum verbo entre outros.

4.2.3.2 Modelo de Espaço do Vector dum Documento

Um Modelo vectorial de Documentos (*A Document Vector Space Model*) pode ser representado num espaço multidimensional em que, cada documento é uma dimensão e cada termo é um vector dentro do espaço de documento.

Também, os termos são as coordenadas e documentos são vectores num espaço de termos. Dois termos são similares se os seus vectores correspondentes estão próximos

entre si. Neste contexto, pode-se calcular a similaridade entre os vectores ***i*** e ***j*** através da seguinte fórmula:

$$\cos(\mathbf{i},\mathbf{j}) = \frac{\mathbf{i} \cdot \mathbf{j}}{\sqrt{\mathbf{i} \cdot \mathbf{i} \times \mathbf{j} \cdot \mathbf{j}}}$$

Em que, o ***cos (i, j)*** permite medir o ângulo entre **i** e **j**, os termos similares tendem a ocorrer nos mesmos documentos e o ângulo entre eles é pequeno. Quando não são similares os termos não ocorrem nos mesmos documentos e o ângulo entre eles é fechado a **zero** (0).

4.2.3.3 Thesaurus com Palavras Infrequentes

Além dos métodos enumerados existe um outro método chamado de **Thesaurus com palavras infrequentes** (*A thesaurus of infrequent words*) onde, o cálculo da similaridade entre termos tem como foco a construção dum *thesaurus* com palavras que repetem poucas vezes num documento de modo a ser possível utilizar essas palavras para a recuperação de documentos.

O método da construção de **Thesaurus com palavras infrequentes,** defende o cálculo do agrupamento de documentos (grupos de documentos) de acordo com a sua similaridade e deste grupo (cluster) seleccionam-se termos discriminadores indiferentes para construir classes do thesaurus.

Um bom discriminador é o termo que tende a aumentar a distância entre documentos e um mau discriminador tende a diminuir a distância entre documentos e um discriminador indiferente não altera a distância entre documentos. Para calcular a discriminação de termos, é muitas vezes utilizado a fórmula do cálculo da frequência dum termo em documentos (número de documentos que um termo aparece) e, quando os termos aparecem menos de 1% nos documentos são discriminadores indiferentes, quando aparecem mais do que 1% e menos do que 10% são bons discriminadores e quando são muito frequentes são maus discriminadores.

Construir um *thesaurus* com termos de baixa frequência para formar classes do referido thesaurus, seria agrupar termos de baixa frequência através dum algoritmo de agrupamento. A similaridade entre agrupamentos (*clusters*) é definida como um mínimo de todas as similaridades calculadas pelo co-seno entre pares de documentos (dois clusters). Depois do cálculo de similaridade entre clusters, os termos de baixa frequência são calculada em cada cluster para formar a corresponde classe do thesaurus.

4.2.4 Representação Textual

4.2.4.1 Sumarização

Sumarização é um processo que produz uma representação sumária dum texto, documento ou grupos de documentos, (Witten 2005), ou seja, permite reduzir o tamanho e o nível de detalhe dum texto, garantindo a essência do seu conteúdo sem perda das palavras-chave e dos objectivos (Fan, Wallace et al. 2005). Por exemplo, as primeiras frases e parágrafos podem constituir num excelente resumo.

É composta por 3 fases, segundo (Lin 2009): **Interpretação** para converter o texto em código de representação, **Transformação** para transformar o texto e procurar frases ou palavras mais importantes e **Geração** para gerar o sumário.

4.2.4.2 Agrupamento

O Agrupamento é uma técnica de processamento textual que permite agrupar um conjunto de documentos desorganizados em grupos similares, (Passarin 2005). Isto quer dizer que se analisam os documentos e criam-se grupos similares de acordo com o conteúdo retratado nos mesmos.

Existem dois tipos de agrupamento: Agrupamento por Partição em que os documentos são distribuídos em grupos distintos, sem nenhuma interligação directa entre os grupos e Agrupamento hierárquico onde é possível a interligação entre os grupos.

No processo de agrupamento, utilizam-se algoritmos como *K-mean*, agrupamento aglomerativo hierárquico, entre outros.

4.2.4.3 Categorização

É uma outra técnica que identifica palavras ou tópicos num texto e faz a sua associação a uma ou mais categorias pré-definidas, (Baas 2008), ou seja, analisa um texto ou um conjunto de documentos e associa-os a uma ou mais categorias levando em consideração os termos existentes nestes documentos.

Para categorizar os documentos utilizam-se classificadores probabilísticos, regressão lógica *Bayesiana*, árvores de decisão e regras de associação, redes neuronais, *support vector machines,* entre outros (Feldman and Sanger 2007).

4.3 Pós-Processamento

O Pós-processamento consiste na avaliação e validação dos resultados obtidos na fase de análise, e tem como principal objectivo melhorar a compreensão do conhecimento descoberto pelo algoritmo utilizado na etapa do processamento, validando-o através de medidas da qualidade.

4.3.1 Critérios para Avaliação da Qualidade

As duas principais variáveis para analisar a qualidade de resultados obtidos são *Precicion* e *Recall*.

Precision é definido como o número de documentos relevantes recuperados numa pesquisa dividido pelo número de documentos recuperados na mesma pesquisa.

$$\boldsymbol{Precision} = \frac{(Documentos\ Relevantes) \cap (Documentos\ Recuperados)}{(Documentos\ Recuperados)}$$

Recall é definido como número de documentos relevantes recuperados numa pesquisa dividido pelo total dos documentos relevantes existente, (Waegel 2006).

$$\boldsymbol{Recall} = \frac{(Documentos\ Relevantes) \cap (documentos\ recuperados)}{documentos\ relevantes}$$

Depois de calculados o *Precision* e o *Recall* calcula-se o *F-measure* que é uma média ponderada ou a combinação entre o *Precision* e o *Recall* através da seguinte fórmula:

$$\boldsymbol{F} = 2.\frac{precision\ .recall}{precision + recall}$$

4.4 Sumário

Neste capítulo, falou-se das várias etapas do processamento textual. Na etapa do pré-processamento abordaram-se técnicas como correcção ortográfica, remoção de *stopwords*, lematização e *n-grans*. Na do processamento falou-se de técnicas de processamento como o cálculo do peso dum termo num documento, extracção de palavras-chave e de frases-chave, descoberta de termos similares através indicação dos vários métodos existentes e a representação textual como a sumarização, o agrupamento e a categorização. Finalmente no pós-processamento falou-se da avaliação da qualidade dos resultados fazendo referência a conceitos como o *Precision, Recall* e *F-Measure*.

Para o caso prático pretende-se aplicar as técnicas de pré-processamento como remoção de *stopwords* e lematização, do processamento como o cálculo da frequência dos termos, a matriz de termos por documentos e a aplicação de algoritmos de classificação e agrupamento e do pós-processamento com o calculo de *precision, recall* e *f-measure*.

Capítulo 5

Ferramentas *Text Mining*

Nos capítulos anteriores falou-se da descoberta do conhecimento em bases de dados (**KDD**) com um enfoque especial para as bases de dados textuais, aliado aos métodos e técnicas para a extracção e descoberta do conhecimento em texto. Para que isto se efective são necessários documentos de texto e ambientes ou ferramentas de desenvolvimento que respeitem o processo de **KDD** e que possibilitem a aplicação dos métodos e técnicas da mineração em texto.

Com este capítulo pretende-se fazer o levantamento de algumas ferramentas de *Text Mining* e a sua caracterização, com um destaque particular para as ferramentas de código aberto (*open source)*, de modo a possibilitar a aplicação das várias técnicas de processamento textual e respectivos algoritmos ao caso de estudo.

5.1 Critérios da Selecção das Ferramentas

Em qualquer processo da descoberta do conhecimento, existem várias etapas do processamento de dados, por conseguinte, quando se escolhe uma ferramenta deve-se respeitar essas etapas.

Pensa-se que escolher ferramentas em mineração de texto é uma etapa muito importante, porque existem muitas ferramentas com características diferenciadas, por isso, é necessário determinar muito bem os objectivos de mineração e escolher ferramentas que possibilitem a materialização desses objectivos.

Para a selecção das ferramentas usou-se, como base, o portal www.kdnuggets.com[1].

Ferramentas	URL
Eaagle	http://www.eaagle.com/
megaputer	http://www.megaputer.com/
vantagepoint	http://www.thevantagepoint.com/
Temis	http://www.temis.com/
SAS Text Miner	http://www.sas.com/technologies/analytics/datamining/textminer/
VantagePoint	http://www.thevantagepoint.com/
Visual Text	http://www.textanalysis.com
WordStat	http://www.provalisresearch.com
GATE	http://gate.ac.uk/
Rapid Miner	http://rapid-i.com/content/view/73/148/
R	http://www.r-project.org/
WeKa/Kea	http://www.cs.waikato.ac.nz/ml/weka/

Tabela 1 – Listagem de Ferramentas *Text Mining*

As ferramentas supra indicadas na tabela 1, representam uma pequena parte das ferramentas de *Text Mining* que existem. Foram escolhidas porque contêm algumas características pretendidas para a materialização do estudo prático.

O processo da escolha da ferramenta a ser utilizada no caso prático não foi fácil, dado que muitas ferramentas não são de utilização livre. Para a avaliação das ferramentas usou-se os critérios propostos por (Cruz 2007) e seleccionou-se alguns itens tais como: Plataforma, ser *open source*, interface gráfica, conectividade a base de texto em diferentes formatos (txt, pdf, Web e email), possibilidade de integração com outras ferramentas de processamento da linguagem natural e capacidade de suportar as etapas da descoberta do conhecimento (KDD). Os resultados desta análise encontram-se na tabela 2 abaixo.

[1] Portal que disponibiliza informações de KDD e *Data Mining* através de casos de estudo, notícias, *links*

Características	Ferramentas								
	eaagle	**megaputer**	**SAS Text Miner**	**Visual Text**	**WordStat**	**GATE**	**Rapid Miner**	**R**	**WeKa/Kea**
Licença	S	S	S	S	S	N	N	N	N
Multi-plataforma	N	N	N	N	N	S	N	S	S
Open Souce	N	N	N	N	N	S	S	S	S
Interface gráfica	S	S	S	S	S	S	S	N	S
Conectividade a Base de Texto	S	S	S	S	S	S	S	S	S
Integração Software PLN[2]	N	N	N	N	N	S	S	S	N
Processo KDD	S	S	S	S	S	S	S	S	S
Algoritmos KDD	N	S	S	S	S	S	S	S	S

Tabela 2 – Caracterização das Ferramentas *Text Mining*

Da análise do quadro, pode-se observar que as ferramentas de código aberto são: GATE, Rapid Miner, R e Weka/Kea, e, por serem de código aberto foram escolhidas para uma caracterização mais aprofundada.

5.2 Caracterização das Ferramentas Seleccionadas

5.2.1 *Rapid Miner*

Rapid Miner, é uma ferramenta *open source*, desenvolvida para *Data Mining* e *Text Mining*. O pacote *Rapid Text Miner* permite fazer o pré-processamento de texto através do processo de divisão do documento num vector de palavras *e* o cálculo da frequência dos termos e dos documentos. Para a aplicação dos métodos de aprendizagem máquina suporta os algoritmos *Naive Bayes, Suport Vector Machines* e, entre outros.

Além destas características esta ferramenta permite a classificação automática, o agrupamento, a análise semântica do texto e a extracção de conhecimento. Em termos de

[2] PLN- Processamento da linguagem Natural

conexão à base de dados permite conexão a estruturas em *xml* e a diferentes sistemas de gestão de bases de dados.

5.2.2 *Weka / Kea*

Weka (*Waikato Environenmenet for Knowledge Analysis*), disponível a partir do sítio http://www.cs.waikato.ac.nz/ml/weka/, é uma ferramenta *Open Source* desenvolvida pela Universidade de Waikato – Nova Zelândia.

Segundo (Baas 2008), *Weka* é uma colecção de algoritmos de Aprendizagem Máquina e de ferramentas de *Data Mining*, implementado em JAVA, que inclui métodos para todos os problemas padrões de *Data Mining*: preparação de dados, pré-processamento, Regressão, Classificação, agrupamento, Regras de associação, Selecção de atributos, entrada de dados e apresentação do resultado da análise.

KEA (*Keyphrase Extraction Algorithm*), disponível a partir do sítio http://www.nzdl.org/Kea/, é também um software código aberto, multi-plataforma (implementado em Java), desenvolvido pela mesma Universidade que funciona como uma extensão do Weka. Esta ferramenta foi desenvolvida para o tratamento de grandes colecções de documentos de texto.

É um algoritmo para a extracção de palavras-chave a partir de documentos de texto que pode ser usado para a indexação livre ou indexação através de um dicionário de palavras. Permite a remoção de *Stopwords*, *Stemmining*, extracção de termos candidatos, cálculo de ocorrência dos termos e extracção de palavras-chave.

Weka/KEA, é uma integração do *KEA* dentro do *Weka*, acumulando assim as características tanto do Weka como KEA e, utilizada em mineração de texto.

5.2.3 *Gate*

General Architecture for Text Engineering (GATE) desenvolvido pela Universidade de *Sheffield* – Inglaterra, é um software livre multi-plataforma (implementado em JAVA) que pode ser usado para desenvolver outras ferramentas de processamento de linguagem natural e extracção de informação.

Suporta funcionalidades como lematização, anotação, extracção de informação em ficheiros de vários formatos, (XML, e-mail, páginas Web, texto e etc). Pode ser integrada com outras ferramentas de processamento da linguagem natural como KEA, WEKA, UIMA, WordNet, ANNIE, *Information Retrieval*, *Machine learning* e entre outras.

5.2.4 *R/tm*

O R (*The R Project for Statistical Computing*), é um *software* livre para a análise estatística e concepção de gráficos que funciona em muitas plataformas. Permite uma grande variedade de análise estatística, classificação e agrupamento. Pode ser estendido com pacotes e para o caso de Text Mining é utilizado um pacote que se chama **tm** (*Text Mining in R)* proposto por (Feinerer, Hornik et al. 2008).

O **R/tm** é uma ferramenta *open source*, composta por um conjunto de bibliotecas específicas, que podem ser utilizadas de acordo com os interesses de cada utilizador. Como é open *source* os pacotes são constantemente actualizados tanto a nível de funcionalidades como de sintaxe. Possibilita também a integração doutras ferramentas como Weka através do Rweka, Kea através do RKea, NLP através de openNLP, entre outras.

Esta ferramenta, permite fazer o pré-processamento com técnicas como remoção de *stopwords* e lematização, permite criar matriz de termos dos corpora de treino e de teste. Além de permitir integração com inúmeras ferramentas de processamento de linguagem natural, *data mining* e de análise estatística, permite fazer a contagem de termos e a

estatística de termos por documentos. Suporta algoritmos como Knn, SVM e knnflex e, entre outros.

5.3 Análise comparativa das Ferramentas

A análise comparativa das ferramentas, faz-se através duma tabela comparativa de características específicas de ferramentas de mineração de texto (Tabela 3) abaixo.

	Ferramentas			
Características	Rapid Miner	Weka/ Kea	Gate	R/tm
Divisão do documento num vector de palavras	**S**			
Stopword				**S**
Stemmining		**S**	**S**	**S**
Anotação			**S**	
Algoritmo Knn (*K - nearest Neighbor*)	**S**	**S**		**S**
Algoritmo NB (*Naive Bayes)*	**S**	**S**		
Algoritmo SVM (Suport *Vector Machines)*	**S**	**S**		**S**
Classificação	**s**	**S**	**S**	**S**
Agrupamento	**S**	**S**		**S**
Classificação Automática	**S**			**S**
Criação de dicionário de palavras (*Thesaurus)*			**S**	
Cálculo da Frequência de termos	**S**			**S**
Cálculo da Frequência dos termos por Documentos	**S**		**S**	**S**
Similaridade dos termos				**S**
Extracção de termos candidatos			**S**	**S**
Extracção de palavras-chave		**S**	**S**	
Extracção de informação	**S**		**S**	**S**
Cálculo de *Precision* e *Recall* automaticamente			**S**	
Integração com KEA		**S**	**S**	**S**
Integração com WordNet			**S**	**S**

Tabela 3 – Comparação das Ferramentas *Text Mining*

Das informações observadas na tabela acima, pode-se concluir que a nível de pré-processamento (Divisão do documento num vector de palavras*, Stopword* e lematização) o R/tm oferece mais opções. Para o processo de anotação e capacidade da criação de dicionário de palavras apenas o GATE permite fazê-los. Em relação aos algoritmos de

descoberta de conhecimento há uma maior frequência para o *K - nearest Neighbor* (Knn) *e o Suport Vector Machines* (SVM) permitido por Rapid Miner, Weka/KEA e R/tm.

A nível de técnicas de processamento como classificação e agrupamento todas têm uma maior inclinação para a classificação. No que respeita a capacidade do cálculo da frequência de termos o R/Tm e RapidMiner oferecem mais opções dado que permitem fazer o cálculo da frequência de termos por Documentos. O R/Tm particularmente permite fazer a análise da similaridade dos termos e a extracção de termos candidatos. Relativamente à extracção de palavras-chave e de informações Weka/KEA e GATE possibilitam a extracção de palavras-chave, enquanto apenas o GATE permite o cálculo automático de *precision* e *recall.*

Para terminar, no que concerne à possibilidade de integrar outras ferramentas o GATE e o R/tm mostram ser mais flexíveis uma vez que permitem a integração o KEA e o WordNet para além de muitas outras ferramentas.

5.4 Experimentação das Ferramentas

Para a concretização deste trabalho testou-se um conjunto de ferramentas *Text Mining,* de utilização livre, uma vez que um dos requisitos colocados à partida foi a opção por ferramentas de utilização livre para a implementação do estudo de caso.

Para isso, optou-se por testar as funcionalidades de todas as ferramentas de utilização livre *(*Weka/Kea, Rapid Miner, Gate e o R/tm*),* referenciadas na secção anterior, para depois escolher a que apresenta melhores facilidades para a materialização do caso prático.

No que diz respeito aos testes com Weka/KEA, experimentaram-se as funcionalidades do *Text Mining* no WEKA verificou-se que na etapa do pré-processamento esta ferramenta não aceita ficheiros com extensão txt e pdf.

Dado que o objectivo era testar o processamento de ficheiros de texto, testou-se a possibilidade de integração do KEA com o WEKA para a extracção de palavras-chave dentro do Texto e, não se obteve sucesso. Fez-se algumas pesquisas e leu-se alguns artigos que faziam referência ao KEA mas, não se conseguiu informações detalhadas que possibilitassem fazer testes.
Para esta ferramenta, pensou-se que os algoritmos que existem no Weka aliados às funcionalidades do algoritmo KEA seria uma óptima escolha mas, como não se conseguiu muitos instrumentos que motivassem a sua exploração, não se optou pela sua utilização.

Com o *Rapid Miner*, depois da sua instalação fez-se a integração do *plugin* para *Text Mining*. Explorou-se funcionalidades como o cálculo da ocorrência e frequência dos termos, remoção de *stopwords*, lematização, conversão de maiúsculas para minúsculas, *n-gram* com um texto em inglês. Durante a sua exploração, observou-se que além de *Data* e *Text Mining,* suporta funcionalidades de *Web Mining,* permite a integração de ferramentas como Weka, wordnet, WVtool e, entre outras, integra vários sistemas de gestão de bases de dados, ferramentas do OLAP, algoritmos de aprendizagem máquina, categorização, clusterização agrupamento e a sumarização do texto.

Ao longo dos testes, sentiu-se algumas dificuldades nomeadamente a integração de ferramentas como WVtool, alteração de algumas parametrizações como a lista de *stopwords* para português a partir da cópia do código para eclipse. Apesar de esta ferramenta gerar ficheiros em formato xml, permitir múltiplas funcionalidades de pré-processamento, integrar vários algoritmos que poderiam ser utilizados, não se conseguiu descobrir uma forma de gerar ficheiros xml de acordo com um formato desejado (tendo em conta os conteúdos que aparecem no texto).
Não se conseguiu também processar vários ficheiros simultaneamente, comparar corpus e aplicar algoritmos aos corpora. Por todas essas limitações, esta ferramenta não figurou uma boa opção, por isso desistiu-se de investir o tempo na sua exploração.

Relativamente ao Gate, primeiramente fez-se a exploração das sessões em vídeo disponíveis on-line para o entendimento do seu funcionamento. Durante a sua exploração observou-se que esta ferramenta trazia várias ferramentas de processamento de texto, em forma de *plugins,* como *KEA*, *WordNet*, *Information Retrieval*, *stemmer Snowball* e entre muitas outras*.*

Fez-se a utilização dos ficheiros em formato txt e xml, para a execução dum conjunto de testes designadamente a anotação dum texto manualmente, cálculo de ocorrência dos termos, *precision*, *recall* e *f-measure,* criação e população dos corpora de treino e teste, alteração da lista de *stopwords*, adaptação da lista de *Gazetteer* (Lista de termos por categoria como abreviações, países, cidades, meses, dias de semana, horas, e etc.) para termos do *eurovoc* em português e, testou-se os softwares integrados como *plugins* aos corpora de treino e de teste.

A maior dificuldade encontrada com esta ferramenta foi que nos testes, particularmente com KEA não se conseguia ver os resultados, ou seja, supostamente a aplicação criava um ficheiro em xml em algum lugar mas, não se conseguiu descobrir onde é que colocava este ficheiro e, consequentemente não se conseguia ter acesso a alguns dados importantes para a interpretação dos resultados obtidos. O facto de não se conseguir visualizar o resultado dos testes não se optou por esta ferramenta embora se tenha investido muito tempo com a sua exploração.

Com o R/tm encontrou-se um ambiente diferente dos outros, uma vez que esta não tem uma interface gráfica. Durante a sua exploração, constatou-se que *tm (Text Mining)* é uma das dezenas de pacotes existentes no R e que o seu funcionamento exigia comandos específicos.

Com esta ferramenta, fez-se a criação dos corpora de treino e teste, aplicou-se as etapas de pré-processamento como conversão de maiúsculas para minúsculas, remoção de *stopwords* e lematização, criaram-se as matrizes dos corpora de treino e de teste e fez-se a aplicação dos algoritmos.

Durante a sua exploração, encontraram-se alguns problemas designadamente a constante mudança das versões que requeriam a adaptação dos *scripts* para cada versão. Para a resolução deste problema concentrou-se numa única versão. Um outro problema encontrado, foi a aplicabilidade dos algoritmos em algumas versões, dado que nem todos foram desenvolvidos para textos. Foi necessário encontrar algoritmos específicos para a classificação de textos, o que implicou testar vários algoritmos suportados pelo R.

Comparando o R/tm com outras ferramentas testadas como o Gate e o RapidMiner, esta ferramenta permite aplicar de forma mais eficiente as técnicas de pré-processamento e análise algorítmica. O Gate apesar de ser uma ferramenta *Text Mining* com uma interface gráfica mais amigável é mais orientada para o processamento da linguagem natural com ficheiros em formato *xml* , o foco neste caso prático, era ficheiros em formato *txt*.

A escolha da ferramenta não foi uma tarefa fácil porque todas faziam referências a aspectos mencionados no estado da arte mas, apresentavam limitações e não se conseguia com nenhuma delas executar uma experiência completa da mineração de texto a nível de pré-processamento (extracção de stopwords, ngram, lematização e etc.), processamento (aplicação de vários algoritmos) e pós-processamento (qualidade dos resultados (*Precision, Recall* e *F-measure*).

Para todas inscreveu-se nos fóruns mas, às vezes demora-se muito tempo para se revolver um problema através do fórum e sentiu-se que estava-se perdendo muito tempo com algumas particularmente (GATE e Rapid Miner). Pensou-se que uma forma de resolver este problema seria a criação duma plataforma de raiz em Java, reutilizando os códigos dado que todas essas ferramentas foram desenvolvidas em Java (Weka, KEA, GATE e RapidMiner) mas, como seria um trabalho que exigia outros caminhos metodológicos da investigação não se optou por esta via dado que o estado da arte já estava quase concluído.

Depois das várias experiências realizadas com as ferramentas *Text Mining* supramencionadas decidiu-se utilizar o **R/tm**, porque ao longo dos testes descobriu-se que com esta ferramenta é possível realizar testes de pré-processamento e aplicação de algoritmos de forma mais directa, por isso, considerou-se que poderia ser ideal para a realização das experiências que se pretendia implementar a nível do pré-processamento e análise de dados.

5.5 Sumário

Neste capítulo, primeiramente fez-se um levantamento de ferramentas *Text Mining* com características para fazer mineração de texto, tanto comerciais como de código aberto. Depois fez-se uma segunda selecção de ferramentas com enfoque específico para ferramentas *open source*, que possibilitem a aplicação das várias técnicas de pré-processamento, aplicação de algoritmos e extracção da informação. Para o caso prático escolheu-se o **R/tm** porque esta ferramenta tem um conjunto de funcionalidades de pré-processamento, análise da frequência dos termos e algoritmos que permitem levar a cabo a experiência de mineração textual. No próximo capítulo vai-se abordar este aspecto com mais detalhe.

Capítulo 6

Estudo de Caso

Pretende-se com este capítulo, fazer a experimentação dos conceitos teóricos explorados nos capítulos anteriores, aplicando-os à problemática da anotação dos Sessões Parlamentares da Assembleia da República Portuguesa. Primeiramente contextualiza-se a problemática estudada, seguidamente caracteriza-se os instrumentos utilizados para a materialização das experimentações a nível das ferramentas, dos textos utilizados nos corpora, dos resultados do pré-processamento e da aplicação e dos algoritmos, finalizando com uma análise da metodologia utilizada e dos resultados conseguidos.

6.1 Contextualização da Problemática do Caso de Estudo

A Assembleia da República Portuguesa tem vindo a fazer um conjunto de investimentos, nos últimos anos, para disponibilizar informações das Actividades Parlamentares através da Internet. O Sistema de Arquivo Áudio Visual e o Projecto dos Debates Digitais desenvolvidos pela Universidade de Aveiro, no âmbito duma parceria entre essas duas instituições, são alguns exemplos destes investimentos.

O Sistema de Arquivo Áudio Visual disponível através do sítio *http://av.parlamento.pt/*, foi desenvolvido para dar continuidade ao Projecto dos Debates Digitais, um outro sistema, que possibilitou o armazenamento dos debates parlamentares e de todos os documentos relacionados com a actividade parlamentar publicados desde a preparação da 1ª Constituição Portuguesa em 1821.

Segundo (Almeida, Martins et al. 2005), o Sistema de Arquivo Áudio Visual é composto por um conjunto de programas que permitem organizar, indexar e anotar o conteúdo dos

segmentos de vídeo através duma estrutura de dados em XML permitindo assim pesquisar recortes audiovisuais (vídeos) das Sessões Parlamentares e doutras actividades Parlamentares.

Segundo o autor supra citado, este sistema permitiu uma maior democratização dos conteúdos da Assembleia da República, tanto para os parlamentares como para o público em geral, uma vez que tornou possível visualizar as intervenções de cada deputado em cada Sessão Parlamentar, através da Internet.

Apesar da excelência do sistema de anotações do Arquivo Áudio Visual, da sua importância para a Assembleia da República na documentação das Sessões Parlamentares e disponibilização das Actividades do Parlamento através da Internet, existem um conjunto de constrangimentos inerentes a esse sistema:

1. A anotação dos arquivos audiovisuais é um processo que decorre 2 ou 3 dias após a realização duma sessão parlamentar. O registo das intervenções em vídeo é feito por especialistas em catalogação e indexação de arquivos através do preenchimento dum número elevado de atributos caracterizando a intervenção feita por um deputado com itens como: nome e categoria do orador ou deputado, duração da intervenção, nome do partido do deputado, transcrição textual com a sua intervenção, palavras-chave, entre outros itens.

2. Tendo em conta que os especialistas em catalogação e indexação de arquivos quando fazem a anotação lêem todo diário duma Sessão Parlamentar e anotam de acordo com as interpretações que fizerem do texto. Isto faz com que a anotação seja um processo moroso.

3. No processo da leitura, se um especialista não interpretar correctamente o conteúdo tratado no texto do diário ou na intervenção dum deputado, pode anotar o texto com palavras-chave incorrectas, porque as palavras-chave

utilizadas na anotação são retiradas duma lista de termos descritores pré-definida, criadas a partir de termos do **eurovoc** [3].

Uma vez que os mesmos debates parlamentares da Assembleia da República estão disponíveis em dois formatos distintos, texto e vídeo, é possível usar os documentos de texto para extrair automaticamente parte da informação necessária para a anotação dos documentos vídeo, permitindo rentabilizar o trabalho realizado pelos técnicos do arquivo (especialistas). Além disso, o processo de anotação contempla apenas a anotação das sessões parlamentares actuais, com início em 2001, pelo que existe um conjunto muito vasto de sessões parlamentares (o arquivo digital da Assembleia da República inclui a transcrição dos debates parlamentares e de todos os documentos publicados desde a preparação da 1ª constituição portuguesa em 1821) que não estão anotadas.

O objectivo deste trabalho consiste em testar a viabilidade da utilização de técnicas de processamento automático de texto para a anotação das sessões dos debates parlamentares da Assembleia da República. Em particular, pretende-se aplicar um conjunto de técnicas de treino e de análise de textos previamente anotados pelos especialistas em catalogação e indexação dos arquivos da AR, para determinar automaticamente quais são os descritores (palavras-chave) que deverão ser associados a cada texto analisado.

Tratando-se de uma primeira abordagem ao problema, o estudo de viabilidade será realizado a dois níveis distintos: associação de descritores por sessão parlamentar e associação de descritores por intervenção dos deputados. Uma sessão parlamentar pode ter a duração de várias horas e inclui a transcrição exacta das intervenções de todos os deputados, que tipicamente têm a duração de alguns minutos cada.

[3] Lista de termos (*Thesaurus*) multilingue que cobre todos os domínios da actividade das comunidades europeias disponível através do sítio http://europa.eu/eurovoc/

O problema que se coloca no primeiro caso será determinar quais os descritores (palavras-chave) associados aos assuntos globais tratados durante uma sessão. No segundo caso, determinar quais os descritores específicos associados à intervenção de um deputado.

Um factor importante que distingue os dois casos e que poderá ter influência decisiva nos resultados será o tamanho dos textos e que será o principal objecto de análise neste trabalho.

6.2 Corpora

6.2.1 Caracterização dos Corpus de Treino e de Teste

Para a realização do estudo foram utilizados os textos do Diário da Assembleia da República retirados a partir da Biblioteca Digital dos Debates Parlamentares, disponíveis a partir do sítio http://debates.parlamento.pt.

Utilizaram-se textos da X (décima) legislatura e de Sessões Legislativas que variam entre Janeiro de 2008 e Setembro de 2009, com o intuito de evitar repetição de um dado documento em vários descritores, o que poderia influenciar o resultado da anotação.

Foram escolhidos 4 descritores, utilizados frequentemente na anotação manual pelos especialistas. A partir desses 4 descritores, fez-se a escolha dos documentos por número do Diário e por mês dentro duma determinada Legislatura e Sessão Legislativa.

Para cada descritor, foram escolhidos dois tipos de documentos:

1. Diários da Assembleia da República referente às Sessões Parlamentares.
2. Intervenção dos deputados numa determinada Sessão Parlamentar escolhida.

O motivo desta preferência é porque quando os especialistas em catalogação e indexação de arquivos fazem uma anotação, seleccionam uma Sessão Parlamentar fazem a anotação manual da Sessão completa e por cada deputado que fez a intervenção na referida Sessão. Esta escolha tem como objectivo comparar o resultado da anotação em textos com tamanhos diferenciados (**Sessões Parlamentares** e **Intervenções dos Deputados**).

Relativamente às características dos dois tipos de documentos pode-se dizer o seguinte:
Para o **Primeiro Caso** – **Anotação das Sessões Parlamentares:**

- O conteúdo é caracterizado por um cabeçalho com a identificação da data (dia de semana, dia, mês e ano), número de série, número da legislatura, identificação da Sessão Legislativa, ano Parlamentar, data da reunião plenária, nome do Presidente e dos Secretários.
- Depois, o Sumário com o assunto tratado e o nome dos deputados que fizeram a intervenção.
- Seguidamente, o texto onde o Presidente declara a abertura da sessão lendo a lista dos deputados presentes na sessão em cada partido político com assento Parlamentar, o texto onde o Presidente manda o secretário fazer a leitura do expediente e o texto onde o Presidente autoriza os deputados a fazerem uma intervenção.
- No final, seguem os textos das intervenções dos deputados e o término da sessão. Para mais informação consultar um exemplo do diário pela referência da capa em anexo (**A1**).

Para o **Segundo Caso** - **Anotação das Intervenções por Deputados em cada Sessão**, os especialistas em catalogação agrupam todas as intervenções feitas por cada deputado e anotam a intervenção do deputado numa sessão parlamentar. O conteúdo caracteriza-se da seguinte forma:

- O extracto do texto onde o presidente de mesa passa a palavra ao deputado para fazer uma Intervenção,
- Este fala, é interpelado por outros deputados, fala novamente até terminar ou o Presidente lhe pede para concluir a intervenção. O deputado conclui a intervenção.
- Depois o presidente pode lhe passar a palavra novamente para responder perguntas ou pedir esclarecimentos.

No que diz respeito ao tamanho dos documentos utilizados nos testes, pode-se dizer que o número de páginas dos Diários utilizados, varia entre 50 a 150 com tamanho variando entre 250K a 700K. Relativamente ao tamanho das intervenções por deputado, varia de meia página a 15 páginas aproximadamente.

6.2.2 Organização do Repositório dos Corpus

Para um tratamento adequado dos textos seleccionados criou-se um repositório com os corpora de treino e de teste de cada um dos descritores escolhidos:

- Para o **1º caso** (sessões parlamentares), foram utilizadas 8 documentos no corpus de Treino e 8 no corpus de Teste.
- Para o **2º caso** (intervenções dos deputados por sessão), foram utilizados 32 documentos para o corpus de Treino e 32 para o corpus de Teste.

No total, foram seleccionados 320 documentos, dado que para a **Sessão Parlamentar** foram seleccionados 64 e para a **Intervenção dos deputados por Sessão** 256. A estrutura do repositório está descrita na ilustração que se segue.

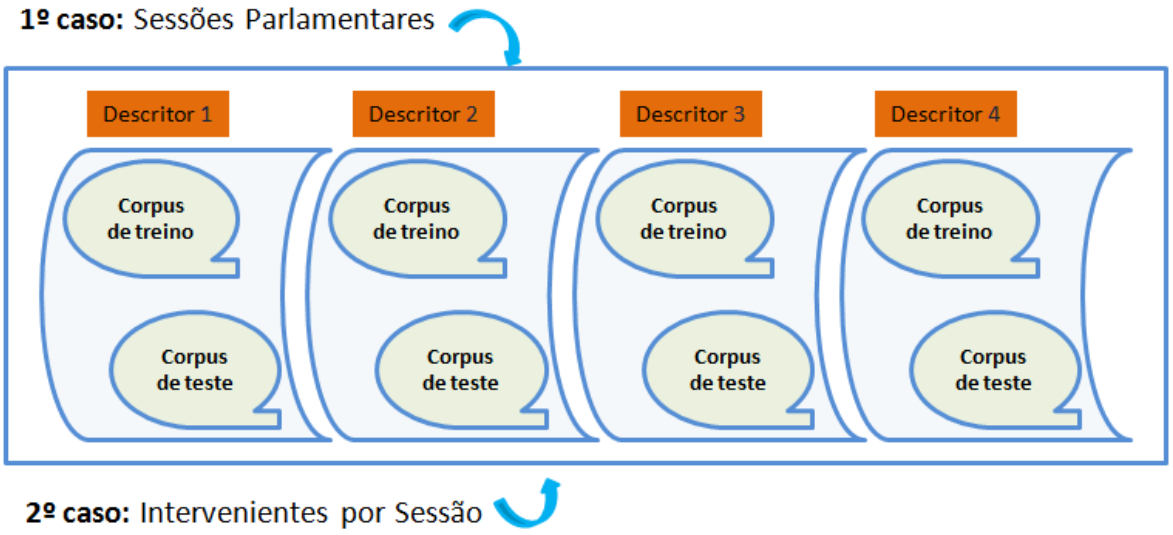


Ilustração 3 - Descrição da Estrutura de agrupamento dos Corpus

A diferença entre os corpora de treino e de teste, é que os de treino têm apenas os documentos que contém os termos descritores que caracterizam o conteúdo retratado nos Diários e nas intervenções específicas de cada deputado, enquanto nos de teste têm documentos que contém os termos descritores em 50% e documentos que falam doutros assuntos em 50%.

Para além da diferença dos termos e supostamente dos conteúdos, existem também documentos de diferentes tamanhos, o que influencia na estatística dos termos com maior frequência.

6.3 Pré-Processamento dos Corpus

O pré-processamento, é uma etapa chave no processamento de texto ou mineração de dados textual, porque se não for bem-feita pode condicionar as etapas subsequentes e a qualidade do resultado textual.

Para este trabalho escolheram-se algumas técnicas de pré-processamento como a remoção de *stopwords* e a lematização com o intento de analisar a influência de cada uma dessas técnicas no resultado final da anotação.

Com o intuito de facilitar o processo de pré-processamento e da aplicação destas técnicas, subdividiu-se o pré-processamento em etapas, segundo a (Ilustração 4) abaixo.

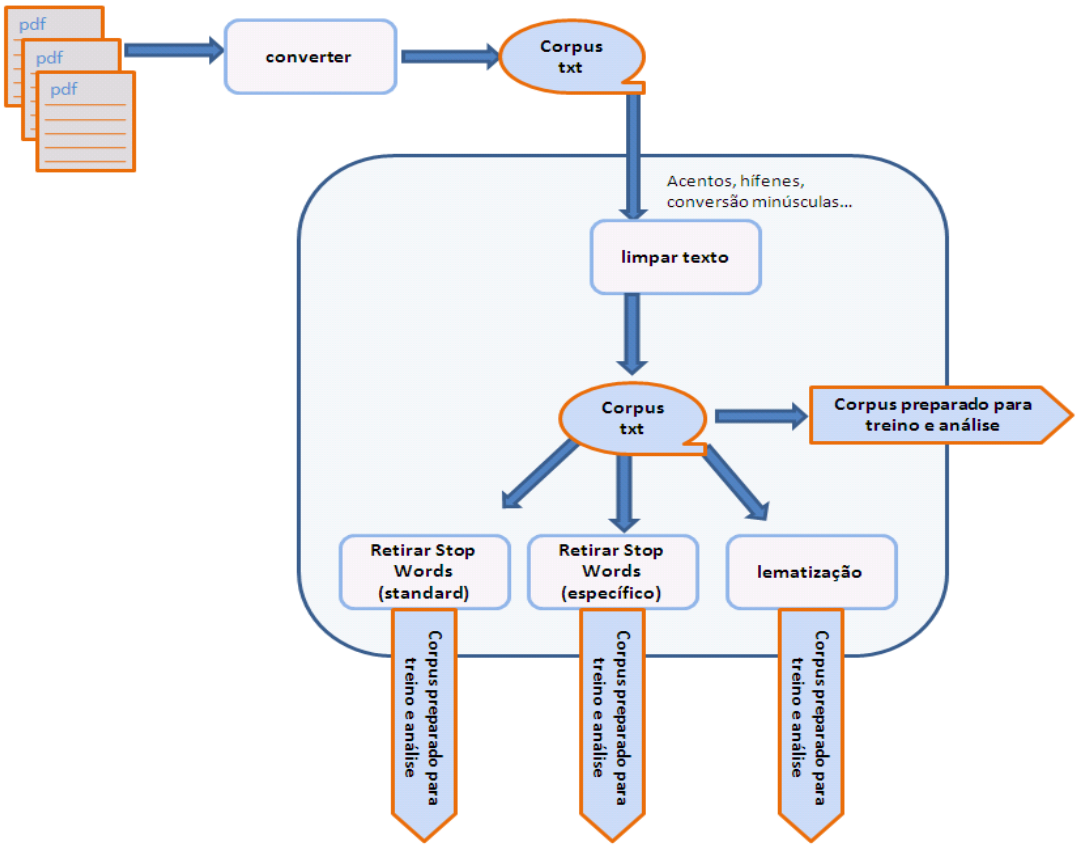


Ilustração 4 - Estrutura do Pré-processamento

Segundo a figura, as sequências do pré-processamento utilizadas tanto para os corpora de teste como para os de treino são:

1. Conversão dos textos.
2. Limpeza do texto.
3. Aplicação das técnicas adicionais do pré-processamento (remoção de *stopword* e/ou lematização).

A etapa da conversão consiste na conversão dos Diários das sessões Parlamentares seleccionadas para o formato *txt* de modo a preparar o texto para a etapa da limpeza. O software utilizado para fazer esta conversão (Pdf em txt) foi o *wordstat*, que é também uma ferramenta de *text mining* com a funcionalidade de conversão.

Um aspecto importante de salientar é que o R permite a utilização dos ficheiros em formato *Pdf* mas, conseguiu-se melhores resultados com ficheiros txt na etapa da limpeza do texto porque os ficheiros Pdf utilizados estão na língua portuguesa contendo caracteres especiais como acentos e o R/tm ainda não permite o processamento destes caracteres especiais alterando as palavras com caracteres que têm acentos para palavras sem significado.
Posto isto, para facilitar o processo da limpeza, converteu-se os ficheiros *pdf* em *txt* e fez-se um programa adicional de remoção de acentos para palavras em português que permite fazer a limpeza sem alterar a estrutura da sintaxe e da semântica do texto.

Depois da conversão dos textos fez-se o agrupamento dos mesmos em corpus de Treino e de Teste referentes a cada descritor seleccionado antes de submetê-los à limpeza.

No processo da limpeza aplicaram-se métodos de remoção de cabeçalhos, números, citações, hífenes, conversão de maiúsculas para maiúsculas e eliminação de espaços em branco com o intuído de preparar o texto para a aplicação de outras técnicas de pré-processamento como a remoção de *stopwords* e lematização.

Depois de submeter os textos ao programa que retira os acentos e à limpeza, considerou-se importante a reaplicação da remoção de pontuação, palavras compostas e eliminação de espaços em branco para eliminar os restantes ruídos que possam estar no texto de modo a complementar o processo de limpeza e assegurar que os corpora fiquem mais leves, além de deixar o texto preparado para a última fase do pré-processamento.

A remoção de *stopwords* e/ou lematização, consiste na etapa final do pré-processamento. No que respeita à remoção de *stopwords,* num primeiro momento utilizou-se a lista de *stopwords* que vem por omissão no **R/tm**, mas depois de vários testes considerou-se que os resultados eram insatisfatórios, por dois motivos:

- Primeiro, porque a lista por defeito não contempla todos os termos considerados *stopwords* em português e além do mais, alguns termos da lista por omissão tinham acentos e estes foram removidos do texto.
- Segundo, porque na mesma lista não existiam termos específicos utilizados diariamente nas Sessões Parlamentares.

Com o intuito de melhorar os resultados fez-se uma actualização na lista de *stopwords* pré-definida, com dois grupos de palavras: Inicialmente, com termos que não constam na lista de *stopwords* por defeito mas, que são considerados *stopwords*[4] em português por alguns investigadores como Rocha, (Rocha 2006). Seguidamente, com termos que se considera *stopwords* no contexto de termos utilizados frequentemente nas sessões parlamentares e que não trazem nenhum acréscimo ao assunto falado numa Sessão Parlamentar como por exemplo: Exactamente, intervenção, deputado (a), governo, presidente, ministro, PS, CDS-PP, PCP, PSD e entre outros termos, ver os anexos (**A2**,**A3** e **A4**).

Depois da aplicação de *stopwords* aplicou-se o método de lematização. Este método permite fazer a uniformização dos termos, tirando os sufixos, que do ponto de vista da contagem dos termos pode ser mais eficiente embora em termos de resultados estatísticos não trouxe grandes melhorias segundo os testes realizados (ver resultado do algoritmo).

[4] http://europa.eu/eurovoc/,

http://www.ranks.nl/stopwords/portugese.html

6.4 Processamento dos Corpus

6.4.1 Processo da Aplicação do Algoritmo

A estrutura utilizada para testar os resultados do algoritmo está descrita na (Ilustração 5).

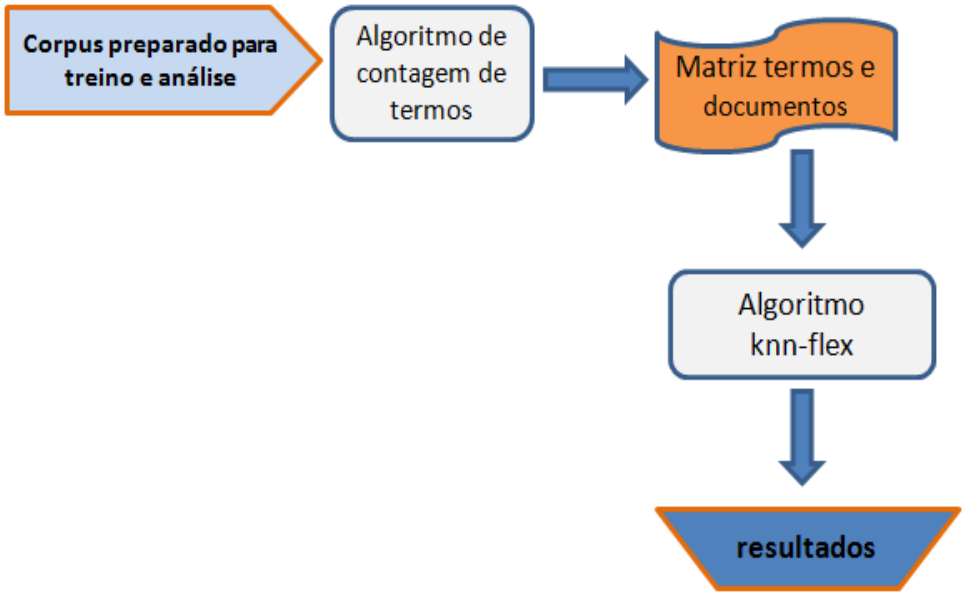


Ilustração 5 – Processo da Análise dos resultados da aplicação do algoritmo *KnnFlex*.

Esta estrutura é constituída por corpora de treino e de teste de cada um dos descritores utilizados, preparados para treino e análise através da aplicação das várias técnicas de pré-processamento.

Pensa-se que a similaridade dos documentos tem uma influência significativa nos resultados dos algoritmos de classificação, porque quanto mais similares forem os documentos melhor é o resultado da classificação.

Para este trabalho aplicaram-se o agrupamento dos documentos similares através do resultado da aplicação dos métodos (*cosine* e *eJaccard*) e a classificação através do algoritmo *KnnFlex.*

A similaridade entre dois documentos representados por vectores de termos com o método *cosine* caracteriza-se pelo cálculo da correlação entre os dois vectores de termos onde o *cosine* é representado pelo ângulo dos dois vectores (Huang 2008). Este autor afirma que o cálculo da similaridade através de *cosine* é independente do tamanho do documento e é uma das mais populares medidas de similaridade aplicada ao texto.

Para o método *eJaccard ou Jaccard* ou Coeficiente de *Jaccard* a similaridade entre dois documentos é feita através da comparação entre o peso da soma dos termos partilhados com a soma do peso dos termos presentes não partilhados em cada um dos dois documentos, (Huang 2008).

Em R, é referenciado no pacote Proxy (**Meyer** and **Buchta** 2010) e utilizados da seguinte forma:

- *dissimilarity (document 1, document 2, method = " **cosine** ")*
- *dissimilarity (document 1, document 2, method = "**eJaccard**")*

Para a classificação, utilizou-se o algoritmo *KnnFlex* que é um algoritmo que surgiu a partir do ***Knn** (K - Nearest Neighbor) ou K* vizinhos mais próximos.
O Knn é um algoritmo de classificação que procura o vizinho mais próximo para classificar, ou seja, decide em que classe coloca um novo caso analisando a distância entre os vizinhos mais próximos, contando o número de caso para cada classe e associando o novo caso à classe semelhante mais próxima. Este algoritmo é mais adequado a dados não estandardizados como texto segundo (Two Crows Corporation 1999).

No calcula de knn é criado um vector de comparação através da utilização da distância euclidiana (Huang 2008) em que, para cada comparação são encontrados os K vizinhos mais próximos. A classificação é feita pela contagem de K maioritário. Se houver ligações para o k mais próximo, então todas as ligações de proximidade são incluídas na contagem.

O algoritmo **KnnFlex** (Brooks 2009) é um *Knn* mais flexível porque permite a execução do algoritmo *Knn* com:

- Especificação de valores para calcular os vizinhos mais próximos,
- Definição de métodos de agregação para sumarizar as respostas através das classes maioritárias,
- Definição do método de manipulação dos resultados da comparação, através da selecção de todas as classes ou então aleatoriamente.

A sua utilização em texto caracteriza-se como um algoritmo de classificação que permite comparar dois corpora (treino e teste) e calcular os Ks vizinhos mais próximos a partir do corpus de treino para o de teste.

Em **R**, a sua utilização é feita através da seguinte forma:
KnnFlex *(train, test, cl, k = 1, prob = FALSE)*, em que:

- train = Matrix do corpus de treino com casos de treino
- test = Matrix do corpus de teste com casos de teste.
- cl = Factor de classificação verdadeira para o corpus de treino
- k = Número de vizinhos considerados
- 1 = Voto mínimo para entrar na lista de candidatos
- prob = Pode ser verdadeira ou falsa. Se for verdadeira, a proporção de votos para a classe que ganhar é retornado como atributo prob = verdadeiro.

Antes da aplicação dos dois métodos foi criada a matriz de termos dos documentos dos corpora (*Term Document Matrix)* através da contagem dos termos de todos os documentos e a criação duma matriz de termos que repetem em todos os documentos.
Utilizando as matrizes de treino e de teste dum determinado descritor, aplicou-se os métodos *cosine* e *ejaccard* para comparar a similaridade dos documentos existentes nos dois corpora. Depois, fez-se o agrupamento dos documentos similares através dum dendograma.

Para o algoritmo *KnnFlex* considerou-se que o corpus de treino tem o papel de treinar o sistema e o de teste de classificar de acordo com os dados treinados. Esses dois corpora foram utilizados através dos dados das matrizes de treino e de teste dum determinado descritor.

6.4.2 Matriz de Termos por Documento

O algoritmo de contagem dos termos permite fazer a contagem do número de vezes que os termos aparecem em cada documento de um corpus. Após a contagem dos termos fez-se a matriz de termos por documento dos corpora de treino e de teste para comparar termos iguais nos dois corpora e consequentemente a sua semelhança.

Abaixo na Tabela 4, Tabela 5 e Tabela 6, pode-se constatar alguns exemplos de matrizes de termos do corpus de treino referente a 1 (um) dos 4 descritores utilizados.

Matriz de Termos

(Sem remoção de *Stopwords*)

Docs	governo	tem	deputado	esta	psd	ministro	mas	muito	porque	uma	nao
1	195	169	143	116	96	293	113	75	70	210	356
2	195	169	143	116	96	293	113	75	70	210	356
3	191	184	105	229	117	87	102	88	66	339	408
4	170	200	128	150	134	46	110	84	71	338	451
5	163	196	140	209	87	27	134	128	79	324	428
6	162	129	90	173	79	20	74	63	40	264	277
7	234	327	211	403	190	26	180	160	126	565	560
8	279	326	283	383	271	69	189	218	123	646	809

Tabela 4 – Matriz de Termos dum corpus sem Remoção de *Stopwords*

Como se pode ver na Tabela 4, uma matriz de termos é constituído por um conjunto de termos que existem em todos os documentos dum corpus e o número de vezes que esses termos aparecem nesses documentos.

Nesta tabela pode-se constatar que existem muitas *stopwords (tem, esta, mas, muito uma e não)* e que assim seria complicado identificar, pelos termos da matriz, o conteúdo dos documentos.

Segundo os testes realizados as matrizes de termos por documentos devolvem melhores resultados se não tiverem palavras consideradas *stopwords* porque quanto mais *stopwords* tiver um documento, maior é a possibilidade de semelhança deste com qualquer outro documento que também tenha *stopwords*.

Na Tabela 5, pode-se constatar que depois de retiradas as *stopwords*, os termos que aparecem na matriz são termos utilizados diariamente no contexto parlamentar.

Matriz de Termos

(Com remoção de *Stopwords* Pré-definida)

Docs \ Terms	socialista	psd	aplausos	ministro	deputado
1	99	193	65	40	175
2	26	98	47	37	79
3	59	142	46	54	144
4	13	84	57	235	119
5	53	119	63	145	130
6	114	173	94	97	231
7	4	82	56	228	98
8	8	28	25	9	10

Tabela 5 – Matriz de Termos dum corpus com Remoção de *Stopwords* Pré-definida

Nesta tabela pode-se deduzir, pelos termos que aparecem na matriz, que o conteúdo do corpus poderia ser algo próximo da política e não da educação, por exemplo. Neste contexto, a incerteza do conteúdo dum corpus pode ser reduzida com a aplicação de *stopwords* especializada.

Matriz de Termos

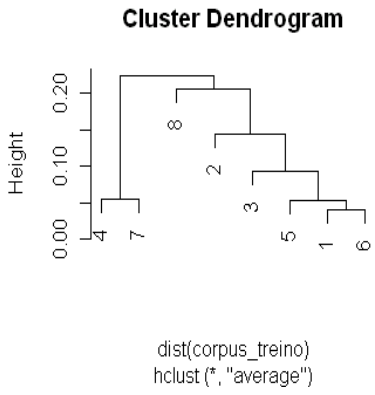
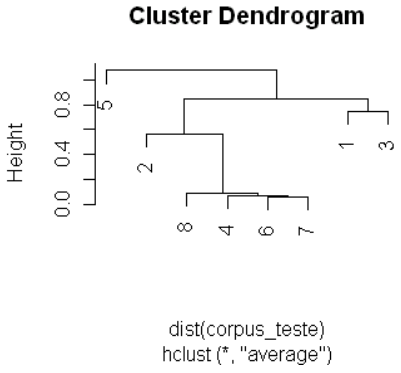
(Com remoção de *Stopwords* Especializadas)

Docs \ Terms	superior	crise	pais	social	lei	escola	abril	ensino	professores
1	4	15	57	67	201	89	30	20	29
2	1	2	44	22	59	138	31	43	131
3	123	37	45	89	56	1	7	126	2
4	4	13	32	19	28	4	4	1	1
5	7	90	103	128	74	26	45	20	10
6	7	21	59	70	134	52	4	11	81
7	17	33	33	81	29	10	22	30	17
8	0	34	43	37	2	2	121	1	0

Tabela 6 – Matriz de Termos dum corpus com Remoção de *Stopwords Especializadas*

Segunda a Tabela 6 pode-se constatar que o conteúdo retratado no corpus, pelos termos que aparecem na matriz, poderia ser algo ligado à educação e isto, mostra claramente que com a aplicação das *stopwords* especializadas o resultado da comparação pode ser mais confiável.

Utilizando a matriz de termos com a remoção de *stopwords* específicas, aplicou-se o método de agrupamento dos corpora onde o resultado é apresentado abaixo através de dendogramas.

Descritor 1 – Corpus de Treino	**Descritor 1** – Corpus de Teste
	

Descritor 1 – Corpus de Treino e Corpus de Teste

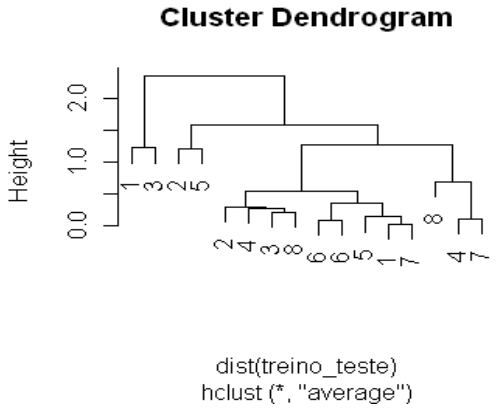


Tabela 7 – Agrupamento dos corpora de teste e de treino

Segundo a os dados dos dendogramas pode-se constatar para o mesmo descritor o seguinte:

No agrupamento do Corpus de Treino, existem 8 documentos agrupados em 6 grupos ou *clusters.* O grupo dos documentos 4 e 7 é diferente do de 1 e 6 e pela distância indicada no cluster entre esses dois grupos pode-se deduzir que os seus conteúdos não são semelhantes. Por outro lado, comparando o grupo dos documentos 1 e 6 com o do 5, pode-se constatar pela distância que estão mais próximos o que significa que os seus conteúdos são mais semelhantes.

No agrupamento do Corpus de Teste existem também 8 documentos agrupados em 6 grupos. Os documentos 8, 4 e 6, apesar de estarem em grupos diferentes, pelas suas proximidades, deduz-se que os seus conteúdos são semelhantes. O documento 5 por estar mais distante destes documentos pressupõe-se que o seu conteúdo não é muito semelhante.

No agrupamento que mistura os corpora de treino e teste, existem 16 documentos e muitos agrupamentos. Nestes agrupamentos nota-se claramente um conjunto maioritário de documentos agrupados em vários grupos com uma distância muito próxima o que mostra claramente que são documentos com conteúdos semelhantes. Como o número total destes documentos ultrapassa 50%, significa que mais de 50% fala sobre os mesmos assuntos.

Fazendo uma análise geral desses resultados, pensa-se que o agrupamento é um método que utilizado em textos sem a remoção das palavras consideradas *stopwords* pode induzir a erros porque as *stopwords* tornam os textos mais semelhantes e em termos gráficos aparecem agrupamentos mais próximos.

6.4.3 Resultados da Anotação com o Algoritmo KnnFlex

6.4.3.1 Influência do Pré-Processamento

A análise do pré-processamento, tem como objectivo analisar o resultado do algoritmo em textos com características diferentes e a viabilidade da aplicação de alguns desses métodos. A metodologia utilizada para avaliar o resultado do algoritmo *KnnFlex*, foi a sua aplicação em textos com e sem *stopwords* e com a lematização.

A Tabela 8 apresenta os resultados da aplicação do *KnnFlex*, para os dois casos (Sessão Parlamentar e Intervenção por deputados):

- Pré-processados sem processamentos adicionais, ou seja, sem remoção de *stopwords* e lematização.
- Sem *stopwords* pré-definidas e específicas, ou seja, removidas as palavras consideradas *stopwords* pré-definidas e específicas.
- Com a aplicação do método da lematização, ou seja, removidas as variações das palavras do mesmo tipo.

	Etapas de Pré-processamento							
	1º Caso: Sessões Parlamentares				**2º Caso: Intervenções por Deputado**			
Documentos	Associados aos Descritores	Acertados	Não Acertados	**% Acertados**	Associados aos Descritores	Acertados	Não Acertados	**% Acertados**
sem processamentos adicionais	32	13	19	41%	32	15	17	47 %
stopwords pré-definidas	32	21	11	66%	32	21	11	66%
stopwords específicas	32	17	15	53%	32	17	16	53%
Aplicação lematização	32	20	12	63%	32	15	17	47%

Tabela 8 – Resultados da aplicação do algoritmo em textos pré-processados

As colunas superiores da tabela descrevem documentos com as seguintes características:

- Associados aos descritores, descreve textos com conteúdos associados aos 4 descritores escolhidos.
- Acertados, representa textos que o algoritmo classificou como textos com conteúdos relacionados ou associados aos descritores escolhidos.
- Não Acertados, representa textos que o algoritmo classificou como textos com outros conteúdos não relacionados com os descritores escolhidos.

Segundo a tabela pode-se constatar que em todas as etapas do pré-processamento foram submetidas ao teste 32 documentos pertencentes aos 4 descritores escolhidos e cada um com 8 documentos.

Fazendo uma análise percentual dos resultados encontrados para o caso das Sessões Parlamentares observou-se que com a aplicação do algoritmo em texto limpo sem outros pré-processamentos adicionais o algoritmo acertou em 41% dos casos. Com a aplicação de *stopwords* pré-definidas (*standard*) constatou-se que o acerto melhorou para 66%. Em *stopwords* especializadas houve uma queda para 53% e com a lematização notou-se um aumentou novamente para 63%.

Para o caso de Intervenções por Deputados observou-se que em texto limpo sem outros pré-processamentos adicionais o algoritmo acertou-se em 47%. Aplicando *stopwords* pré-definidas aos mesmos textos o acerto melhorou para 66%, com *stopwords* especializadas diminuiu para 53%. Com a aplicação da lematização verificou-se que desceu para 47%.

Fazendo uma análise geral segundo os resultados, constatou-se que a aplicação das *stopwords* especializadas não trouxe nenhuma melhoria em relação ao *stopword* pré-definida. Pensa-se que este resultado é consequência da remoção de muitas palavras do contexto parlamentar que se repetem variadas vezes e que com a sua remoção os textos de (treino, teste) passam a ter menos similaridade o que reflecte directamente no resultado da classificação. Isto poderá ser um factor interessante de análise dado que poderia servir para encontrar textos com conteúdos similares sem *stopwords* pré-definidas e nem especializadas.

Também constatou-se que a lematização não trouxe melhorias, por conseguinte, não se justifica a sua aplicação dado que para os dois casos obteve-se melhores resultados em textos com a aplicação de *stopwords* pré-definidas.

Pensa-se que o tamanho dos documentos pode influenciar esses resultados, porque quanto mais extensa for um documento maior é a possibilidade deste ter palavras iguais comparado com outros.

O gráfico abaixo mostra a **influência do tamanho dos documentos** e a variação dos resultados em todas as etapas da análise.

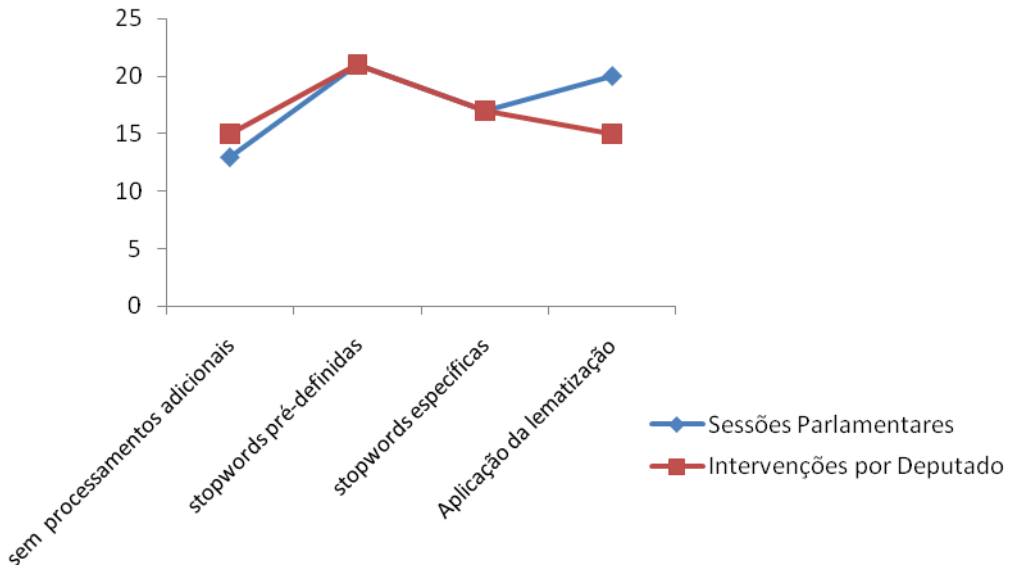


Gráfico 1 – Influência do tamanho dos Documentos no Pré-Processamento

Segundo o gráfico em textos sem processamentos adicionais houve melhores resultados em Intervenções por deputado. Para textos sem *stopwords* pré-definidas e específicas os resultados foram iguais. Com a aplicação da lematização, intervenções por deputado teve melhores resultados.

Fazendo uma apreciação global dos resultados, pode-se constatar que em Sessões Parlamentares registou-se melhores resultados do que em Intervenções por Deputado. Isto significa que o tamanho tem influência no resultado e consequentemente no número de acertos, porque os textos das sessões parlamentares são mais extensos e os conteúdos são mais diversificados (muitos termos).

6.4.3.2 Influência do Tamanho do Corpus de Treino

Para a análise da influência do tamanho dos corpora foram testados corpora com tamanhos diferentes tanto para os textos das Sessões Parlamentares como para os das Intervenções por Deputado. Para o 1º caso foram testadas agrupamentos de 1, 2, 4 até 8 e para o 2º caso agrupamentos de 1,2,4,8,16 até 32.

Foram escolhidos textos submetidos à aplicação de *stopwords* pré-definidas, porque encontrou-se melhores para este tipo de texto na análise do pré-processamento.

	Tamanho do Corpus de Treino							
	1º Caso: Sessões Parlamentares				**2º Caso: Intervenções por Deputado**			
Número de Documentos	Associados aos Descritores	Acertados	Não Acertados	**% Acertados**	Associados aos Descritores	Acertados	Não Acertados	**% Acertados**
1	4	2	2	50%	4	1	1	25%
2	8	7	1	88%	8	2	6	25%
4	16	8	8	50%	16	6	10	38%
8	32	21	11	66%	32	21	11	66%
16					64	21	40	38%
32					128	71	57	55%

Tabela 9 - Resultado da análise do tamanho dos corpora

Fazendo uma análise percentual dos resultados da Tabela 9 pode-se observar que para corpora com 1 documento o algoritmo acertou em 50% para Sessões Parlamentares e 25% para intervenções por deputado.

Para 2 documentos em Sessões Parlamentares subiu para 88% enquanto para intervenções por deputado manteve-se os 25%.

Para 4 em sessões parlamentares desceu para 50% enquanto para intervenções por deputados subiu para 38%. Para 8 documentos o resultado foi exactamente igual, para ambos os casos, dado que os dois tiveram 66% de acertos.

Para 16 a 32 analisados apenas em intervenções por deputados observou-se que com 16 o número de acertos piorou para 38% e com 32 melhorou para 55%.

Relativamente à influência do número dos documentos por corpora constatou-se que em Sessões Parlamentares há maior número de acertos do que em Intervenções por Deputados. Este resultado é ilustrado no gráfico abaixo.

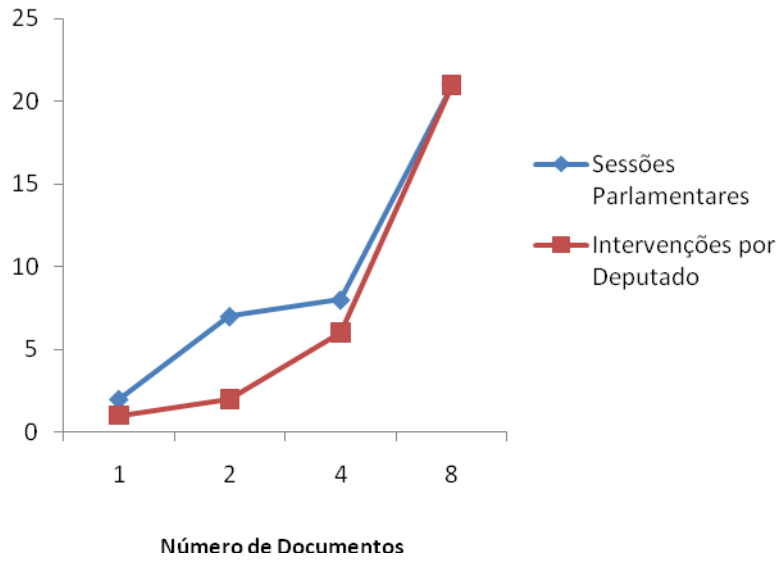


Gráfico 2 - Influência do Tamanho dos Documentos

Segundos os dados apresentados, pode-se constatar que o número ideal de documentos por corpora seria 8 uma vez que para os dois casos encontrou-se melhores resultados com este número.

6.4.3.3 Influência do Tamanho dos Corpora de Treino e de Teste

Para a análise da influência do entrançamento dos textos das sessões parlamentares com intervenções por deputado testou-se o resultado do algoritmo utilizando textos das sessões parlamentares como corpus de treino e intervenções por deputado como corpus de teste e, vice-versa.

Analisando os resultados segundo os dados da Tabela 10, observou-se que com Treino (**sessões**) versus Teste (**intervenções**) o número de acertos foi de 56% e para o caso inverso, Treino (**intervenções**) versus Teste (**sessões**) o algoritmo acertou-se em 53%.

Número de Documentos	SP - Sessões Parlamentares (treino) versus IP - Intervenções por Deputado (teste) e Vive Versa							
	1º Caso: SP - Sessões Parlamentares (treino) versus IP - Intervenções por Deputado (teste)				**2º Caso:** SP - Sessões Parlamentares (teste) versus IP - Intervenções por Deputado (treino)			
	Associados aos Descritores	Acertados	Não Acertados	**% Acertados**	Associados aos Descritores	Acertados	Não Acertados	**% Acertados**
SP versus **IP**	32	18	14	56%				
IP versus **SP**					32	17	15	53%

Tabela 10 - Intervenções versus Sessões e vice-versa

Fazendo uma análise comparativa dos resultados pode-se dizer que para o primeiro caso Treino (**sessões**) versus Teste (**intervenções**) conseguiu-se melhores resultados uma vez que o número de acertos foi de 56% para documentos associados aos descritores.

6.5 Pós-Processamento

Depois das etapas de pré-processamento e processamento faz-se a análise dos resultados através do *Precision*, *Recall* e *F-measeure*. O Resultado é apresentado na Tabela 11.

Qualidade dos Resultados	1º Caso: Sessões Parlamentares			2º Caso: Intervenções por Deputado		
	Precision	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
Sem processamentos adicionais	62%	49%	49%	52%	47%	49%
stopwords pré-definidas	55%	66%	60%	50%	66%	57%
stopwords específicas	49%	53%	51%	52%	53%	32%
Lematização	47%	63	53%	47%	47%	47%

Tabela 11 – Qualidade dos Resultados do Pré-processamento

Analisando os resultados do *f-measure* pode-se observar que para textos limpos sem processamentos adicionais obteve-se melhores resultados em intervenções por deputado e pensa-se que este resultado pode ser a influência do tamanho. Para os outros casos, observou-se que em Sessões parlamentares obteve-se melhores resultados.

Para o caso de *stopwords* específicas os resultados foram uma surpresa porque considerou-se inicialmente que a eliminação de palavras consideradas *stopwords* no contexto dos debates parlamentares melhoraria os resultados, dado que, essas palavras podem ser consideradas ruídos dentro do texto.

Finalmente, fazendo uma apreciação dos resultados alcançados com a lematização considera-se que é inútil a sua aplicação uma vez que não trouxe melhorias para a maioria dos cenários analisados.

6.6 Avaliação das Metodologias e Resultados

Analisando as metodologias pode-se dizer que a escolha dos textos a analisar é muito importante tendo em conta que num primeiro momento se obtiveram resultados insatisfatórios porque a metodologia utilizada para escolher textos foi a pesquisa livre dum determinado termo descritor (por exemplo educação) em documentos pdf. Dado que um documento pode estar relacionado com vários descritores pode acontecer repetição de textos, por isso, optou-se pela escolha de Sessões num determinado mês e numa determinada Legislatura.

Do ponto de vista da eficiência e eficácia do pré-processamento não se aconselha a utilização da lematização no início do pré-processamento, porque da análise de alguns testes realizados este processo remove os sufixos dos termos e desestabiliza o processo da remoção de *stopwords* e consequentemente o resultado do pré-processamento e da aplicação de algoritmos.

Analisando os resultados observou-se que a nível de pré-processamento obteve-se melhores resultados com a aplicação de *stopwords* pré-definidas, embora à partida se pensasse que os resultados obtidos com remoção de *stopwords* específicas seriam mais fiáveis, porque seria possível remover palavras de uso frequente no contexto parlamentar facilitando a extracção de termos que podem ser efectivamente relevantes para anotar os conteúdos retratados no texto.

Analisando os resultados duma forma global verificou-se melhores resultados com os textos das Sessões Parlamentares e pensa-se que este resultado é influenciado pelo tamanho dos textos.

A nível da aplicação do algoritmo *KnnFlex* verificou-se que o nível médio de acertos foi de 50 a 60 %, uma percentagem relativamente baixa mas, tendo em conta que os corpora de testes tiveram 50% dos documentos com assuntos relacionados com os descritores e outros 50% outros assuntos pensa-se o que este resultado é aceitável.

6.7 Sumário

Neste capítulo, fez-se a demonstração das experimentações feitas no caso prático. Falou-se da estrutura do repositório de textos utilizada nos testes e dos resultados da experimentação a nível do pré-processamento, processamento e pós-processamento. No pré-processamento fez-se referência aos resultados da aplicação das técnicas da remoção de *stopwords* e lematização. No processamento apresentou-se os resultados do agrupamento e classificação dos textos e, no pós-processamento fez-se uma análise da qualidade dos resultados encontrados.

Capítulo 7

Conclusões e Perspectivas

Este capítulo apresenta as principais conclusões da dissertação. Primeiramente faz-se uma síntese dos conteúdos teóricos explorados, seguidamente analisa-se os resultados conseguidos e finaliza-se com algumas contribuições e trabalhos futuros.

O trabalho iniciou-se com uma pesquisa sobre o conceito da descoberta do conhecimento em bases de dados, onde se fez um estudo detalhado das várias tecnologias de *Mining* aplicadas aos vários formatos de dados, a arquitectura dum sistema *Text Mining* e as técnicas de Pré-processamento, processamento e pós-processamento de textos. Seguidamente fez-se um estudo comparativo das ferramentas *Text Mining* de utilização livre que poderiam servir para a experimentação dos conceitos teóricos explorados e as suas aplicabilidades à materialização do Caso de Estudo.

Pretendeu-se com este trabalho fazer um estudo de viabilidade da utilização de técnicas de processamento automático de texto para a anotação das sessões dos debates parlamentares da Assembleia da República Portuguesa, previamente anotados pelos especialistas em catalogação e indexação dos arquivos.
Um factor importante que se queria estudar foi se a influência do tamanho dos textos poderia influenciar os resultados. Por conseguinte, o estudo foi realizado a dois níveis: associação de descritores por Sessão Parlamentar e associação de descritores por Intervenção dos Deputados.

Para a sua materialização utilizou-se um conjunto técnicas de pré-processamento, processamento e pós processamento de texto. Primeiramente fez-se a recolha e tratamento de 320 textos, sendo 64 textos dos Diários das Sessões Parlamentares e 256

textos das Intervenções por Deputados. Criou-se duas listas de *stopwords*: Uma com termos específicos do contexto parlamentar outra com um número considerável de termos considerado *stopwords* em Português. Depois, utilizou-se o algoritmo *KnnfLex* para analisar 3 aspectos: A influência do pré-processamento, do tamanho dos documentos e dos corpora de treino e teste.

Na análise da influência do pré-processamento observou-se que a lista de *stopwords* específicas e a lematização não trouxeram melhorias na anotação dado que para os dois casos conseguiu-me melhores resultados com a aplicação de *stopwords* pré-definidas. Relativamente ao tamanho dos documentos, pode-se dizer que o tamanho influência no resultado porque registou-se melhores resultados em Sessões Parlamentares tanto na análise do tamanho dos documentos como na análise da influência dos corpora de treino e de teste. Pensa-se que pelo facto dos textos das sessões parlamentares serem mais extensos e mais heterogéneos, determinou este resultado.

Um outro aspecto importante que se queria estudar era que mecanismos devem ser utilizados para determinar automaticamente quais são os descritores (palavras-chave) que devem ser associados a cada texto analisado. Para este caso, observou-se que a remoção de *stopwords* específicas permite remover palavras de uso frequente num contexto específico possibilitando assim a extracção de termos que podem ser efectivamente relevantes para anotar o conteúdo retratado no texto.

Importa também referir que durante a concepção deste trabalho sentiram-se algumas limitações particularmente dificuldades relacionadas com a parametrização das ferramentas o que poderá ter influenciado os resultados apresentados no caso prático tendo em conta que poderiam ser testados outras técnicas de análise de texto, outros algoritmos de classificação e agrupamento.

Em termos de contribuições, o levantamento e a análise das ferramentas de utilização livre a nível de *text mining*, a lista de *stopwords* específicas e os resultados encontrados poderão servir como pontos de partida para trabalhos de investigação futuros.

Pensa-se que os resultados apresentados poderiam ser melhorados, com uma exploração mais aprofundada das potencialidades das *stopwords* específicas, dado este método permite remover termos típicos utilizados num contexto específicos e que não são considerados relevantes para anotar, determinar ou associar o conteúdo retratado num texto especificamente.

Bibliografia

Almeida, P., M. Fernades, et al. (2005). DisQS - Distributed Query System Based on Web Services for Digital Libraries. ITA 2005 - International Conference on Internet Technologies and Applications, Universidade de Wales,North East Wales, UK, 7-9 September 2005.

Almeida, P., J. A. Martins, et al. (2005). Portuguese Parliamentary AudioVisual Archive: Organizing and Browsing a Mutlimedia Database Fourth International Workshop on Content-Based Multimedia Indexing, IEEE/EURASIP/COST292, Tampere University of Technology, Riga, Latvia, 21-23 June 2005.

Aranha, C. and E. Passos (2006). A Tecnologia de Mineração de Textos. RESI-Revista Electrônica de Sistemas de Informação, Nº2-2006. Lab.ICA Elétrica PUC-Rio, [Agosto 2008], http://revistas.facecla.com.br/index.php/reinfo/article/viewFile/171/66

Baas, T. A. (2008). Data Mining and Text Mining in Business Intelligence and a Preliminary Approach to Combine Data Mining and Text Mining. Department of Computer Science Amsterdam, Vrije Universiteit - Amsterdam. ***PHD*****:** 146.

Brooks, A. D. (2009). Package 'knnflex' - A more flexible KNN, The R Project for Statistical Computing.

Cruz, A. J. R. d. (2007). Data Mining via Redes Neuronais Artificiais e Máquinas de Vectores de Suporte. Escola de Engenharia - Departamento de Sistemas de Informação. Braga, Universidade do Minho.

Elberrichi, Z., Computer Science Department, College of Engineering - University of Sidi Bel-Abbes, Algeria and B. Aljohar, Computer Science & IT college - King Faisal University,Al-Hasa, Saudi Arabia (2007). "N-grams in Texts Categorization." Scientific Journal of King Faisal University (Basic and Applied Sciences) **Vol. 8 No. 2 1428H (2007)**.

Fan, W., L. Wallace, Department of Accounting and Information Systems - Virginia Polytechnic Institute and State University, et al. (2005). Tapping into the Power of Text Mining, Communications of ACM**:** [Junho 2008] http://filebox.vt.edu/users/wfan/paper/text_mining_final_preprint.pdf.

Feinerer, I., K. Hornik, et al. (2008). Text Mining Infrastructure in R. Wirtschaftsuniversitat Wien - Journal of Statistical Software **Volume 25**.

Feldman, R., Bar-Ilan University, Israel and J. Sanger, ABS Ventures, Waltham, Massachusetts (2007). The Text Mining HandBook Advanced Approaches in Analyzing Unstructured Data, Cambridge University.

Gonçalves, a., f. beppler, et al. (2005). UM MODELO BASEADO EM MINERAÇÃO DE TEXTOS VOLTADO A APLICAÇÕES DE GESTÃO DO CONHECIMENTO. O diálogo Universidade-Empresa na Sociedade do Conhecimento, 11., 2005, . K. 2005. SÃO PAULO - ANAIS, Instituto Stela.

Han, J. and M. Kamber (2006). Data Mining Concepts and Techniques. M. R. Jim Gray, Morgan Kaufmann. **second edition**.

Hearst, M. (2003). "What Is Text Mining?" **October 17, 2003**: [Feveiro 2008], http://people.ischool.berkeley.edu/~hearst/text-mining.html.

Hotho, A., KDE Group - University of Kassel, A. N¨urnberger, Information Retrieval Group - School of Computer Science - Otto-von-Guericke-University Magdeburg, et al. (2005). A Brief Survey of Text Mining. [Fevereiro 2008] http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf.

Huang, A. (2008). Similarity Measures for Text Document Clustering. Computer Science Research Student Conference 2008, New Zealand, Department of Computer Science - The University of Waikato, Hamilton, New Zealand.

Khrouf, K. and C. Soulé-Dupuy (2004). A Textual Warehouse Approach:A Web Data Repository, University of Toulouse III. Intelling Agents for Data Mining and Information Retrieval. University of Canberra, Australia, Idea Group Publishing, British Library.

Lin, J. (2009). Summarization. Encyclopedia of Database Systems. Springe, University of Maryland, College Park.

Magalhães, L. H. d. (2008). UMA ANÁLISE DE FERRAMENTAS PARA MINERAÇÃO DE CONTEÚDO DE PÁGINAS WEB. Retrieved [Dezembro 2009], from http://wwwp.coc.ufrj.br/teses/mestrado/Novas_2008/teses/MAGALHAES_LH_08_t_M_int.pdf .

Maimon, O. and L. Rokach (2005). Data Mining and Knowledge Discovery Handbook. Univercity of Israel, Springer.

Medelyan, O. and H. I. Witten (2006). Thesaurus Based Automatic Keyphrase Indexing. Department of Computer Science - University of Waikato, New Zealand, ACM.

Medelyne, O. (2005). Automatic Keyphrase Indexing with a Domain-specific Thesausus. Albert-Ludwigs-University -Ukraine. **Master of Computer Science:** 99.

Meyer, D. and **C. Buchta** (2010). Package 'proxy' - Distance and Similarity Measures. The R Project for Statistical Computing.

Monteiro, L. d. O., I. R. Gomes, et al. (2006). Etapas do Processo de Mineração de Textos – uma abordagem aplicada a textos em Português do Brasil. Anais do XXVI Congresso da SBC - I Workshop de Computação e Aplicações, Campo Grande, Centro Universitário do Pará (CESUPA)

Passarin, D. (2005). Text Mining no Aperfeiçoamento de Consultas e Definição de Contextos de uma Central de Notícias Baseada em RSS. Centro Universitário Luterano de Palmas - Brasil.

Penteado, R. and E. Boutin (2008). Creating Strategic Information for Organizations with Structured Text. Emerging Technologies of Text Mining:Techniques and Applications. Catholic University of Brasilia - Embrapa Food Technology, Brazil, IGI Global.

Pouliquen, B., R. Steinberger, et al. (2003). Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. Ontologies and Information Extraction. Workshop at EUROLAN'2003 - The Semantic Web and Language Technology - Its Potential and Practicalities, Bucharest, 28 July - 8 August 2003, . .

Rocha, P. A. (2006). "CHAVE Collection." Retrieved 28 de Outubro de 2009, from http://linguateca.di.uminho.pt/Paulo/.

Rose, Ø. (2007). Text Mining in Health Records - Classification of Text to Facilitate Information Flow and Data Overview. Department of Computer and Information Science, Norwegian University of Science and Technology. **Master of Computer Science:** 144.

Samat, N. A., "Faculty of Computer Science and Information Technology Faculty of Computer Science and Information Technology University Putra Malaysia, 43400 Serdang, Selangor, Malaysia", M. A. A. Murad, et al. (2008). Term Weighting Schemes Experiment Based on SVD for Malay Text Retrieval. **VOL.8 No.10, October 2008**: [Março 2008] http://paper.ijcsns.org/07_book/200810/20081055.pdf.

Seifert, J. W. (2004). Data Mining: an Overview, CRS IJCSNS International Journal of Computer Science and Network Security Report for Congress - Received through the CRS Web, Analyst in Information Science and Technology Policy - Resources, Science and Industry Division.[Janeiro 2008], http://www.fas.org/irp/crs/RL31798.pdf

Senellart, P. P. and V. D. Blondel (2008). Automatic discovery of similar words. Survey of Text Mining II - Clustering, Classification and Retrieval, Springer**:** [Março 2008] http://www.inma.ucl.ac.be/~blondel/publications/02SB.pdf.

Shetty, S. D. o. C. A., NMAM Institute Of Technology, Nitte,Udupi, Karnataka, India and K. K. D. o. S. Achary, Mangalore University, Mangalagangothri, India (2008). Audio Data Mining Using Multi-perceptron Artificial Neural Network. IJCSNS International Journal of Computer Science and Network Security **VOL.8 No.10, October 2008**.

Simoff, S. J. F. o. I. T., University of Technology, Sydney,, C. I. Djeraba, Nantes University, et al. (2002). MDM/KDD2002: Multimedia Data Mining between Promises and Problems. SIGKDD Explorations **Volume 4, Issue 2**: [Dezembro 2008] http://www.sigkdd.org/explorations/issue4-2/simoff.pdf.

Solka, J. L. (2007). Text Data Mining: Theory and Methods. Naval Surface Warfare Center Dahlgren Division **Statistics Surveys, Volume 2 (2008), 94-112**.

Two Crows Corporation (1999). Introduction to Data Mining and Knowledge Discovery.

Von, H., A. Mehler, et al. (2005). Text Mining. Journal for Computational Linguistics and Language Technology.

Waegel, D. (2006). "The Development of Text-Mining Tools and Algorithms."

Witten, I. H., O. Medelyan, et al. (2006). Finding documents and reading them: Semantic metadata extraction, topic browsing and realistic books. Department of Computer Science, University of Waikato, New Zealand, Proceedings of the 8th Russian Conference on Digital Libraries RCDL'2006, Suzdal, Russia, 2006.

Witten, I. H. C. S., University of Waikato, Hamilton, New Zealand (2005). Text mining. In The Practical Handbook of Internet Computing, Boca Raton, Florida. , 14, 1-22. Chapman & Hall/CRC: [Março 2008] http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf.

Yang, H. (2005). Methodologies for information source selection under distributed information environments. Scholl of Information Technology and Computer Science University of Wollongong. **Phd of Computer Science:** 221.

Zaiane, O. R. (1999). Introduction to Data Mining. CMPUT 690: Principles of Knowledge Discovery in Databases, **University of Alberta, Department of Computing Science**, [setembro 2008] http://www.exinfm.com/pdffiles/intro_dm.pdf.

Anexos

A.1 Capa da Sessão Parlamentar do dia 5 de Ferreiro de 2006

Quinta-feira, 5 de Fevereiro de 2009 I Série — Número 42

DIÁRIO
da Assembleia da República

X LEGISLATURA 4.ª SESSÃO LEGISLATIVA (2008-2009)

REUNIÃO PLENÁRIA DE 4 DE FEVEREIRO DE 2009

Presidente: Ex.mo Sr. Jaime José Matos da Gama

Secretários: Ex.mas Srs. Maria Celeste Lopes da Silva Correia
Fernando Santos Pereira
Rosa Maria da Silva Bastos da Horta Albernaz
Maria Ofélia Fernandes dos Santos Moleiro

SUMÁRIO

O Sr. Presidente declarou aberta a sessão às 15 horas e 10 minutos.

Deu-se conta da entrada na Mesa da proposta de lei n.º 249/X (4.ª), do projecto de resolução n.º 422/X (4.ª) e da interpelação n.º 26/X (4.ª).

A Câmara aprovou um parecer da Comissão de Ética, Sociedade e Cultura relativo à reforma de mandato de uma Deputada do PS.

Procedeu-se ao debate de interpelação n.º 25/X (4.ª) — Sobre a situação social, desemprego e pobreza (PCP), tendo proferido intervenções, na fase de abertura, o Sr. Deputado Bernardino Soares (PCP) e o Sr. Ministro do Trabalho e da Solidariedade Social (Vieira da Silva). Além destes oradores, usaram da palavra durante o debate, e diverso título, os Srs. Deputados Adão Silva (PSD), Mariana Aiveca (BE), Pedro Mota Soares (CDS-PP), Heloísa Apolónia (Os Verdes), Isabel Santos (PS), Jorge Machado (PCP), Miguel Santos (PSD), Ana Drago (BE), Bruno Dias, José Alberto Lourenço, João Oliveira e Agostinho Lopes (PCP), Sónia Fertuzinhos (PS) e Arménio Santos (PSD).

No encerramento do debate proferiram intervenções o Sr. Deputado António Filipe (PCP) e o Sr. Ministro dos Assuntos Parlamentares (Augusto Santos Silva).

O Sr. Presidente encerrou a sessão eram 17 horas e 40 minutos.

A.2 Lista de *Stopwords* Pré-definida

a
à
acerca
agora
algmas
alguns
ali
ambos
antes
ao
aos
apontar
aquela
aquelas
aquele
aqueles
aqui
aquilo
as
às
até
atrás
bem
bom
cada
caminho
cima
com
como
comprido
conhecido
corrente
da
das
de
debaixo
dela
delas
dele
deles
dentro
depois
desde
desligado
deve
devem
deverá
direita
diz
dizer
do
dois
e
é
ela
elas
ele
eles
em
enquanto
então
entre
era
eram
éramos
essa
essas
esse
esses
esta
está
estado
estamos
estão
estar
estará
estas
estava
estavam
estávamos
este
esteja
estejam
estejamos
estes
esteve
estive
estivemos
estiver
estivera
estiveram
estivéramos
estiverem
estivermos
estivesse
estivessem
estivéssemos
estou
eu
fará
faz
fazer
fazia
fez
fim
foi
fomos
for
fora
foram
fôramos
forem
formos
fosse
fossem
fôssemos
fui
há
haja
hajam
hajamos
hão
havemos
hei
horas
houve
houvemos
houver
houvera
houverá
houveram
houvéramos
houverão
houverei
houverem
houveremos
houveria
houveriam
houveríamos
houvermos
houvesse
houvessem
houvéssemos
iniciar
inicio
ir
irá
isso
isto
já
lhe
lhes
ligado
maioria
maiorias
mais
mas
me
mesmo
meu
meus
minha
minhas
muito
muitos
na
não
nas
nem
no
nome
nos
nós
nossa
nossas
nosso
nossos
novo
num
numa
o
onde
os
ou
outro
para
parte
pegar
pela
pelas
pelo
pelos
pessoas
pode
poderá
podia
por
porque
povo
primeiro
qual
qualquer
quando
que
quê
quem
quieto
saber
são
s
seja
sejam
sejamos
sem
ser
será
serão
serei
seremos
seria
seriam
seríamos
seu
seus
só
somente
somos
sou
sua
suas
tal
também
te
tem
tém
têm
temos
tempo
tenha
tenham
tenhamos
tenho
tentar
tentaram
tente
tentei
terá
terão
terei
teremos
teria
teriam
teríamos
teu
teus
teve
tinha
tinham
tínhamos
tipo
tive
tivemos
tiver
tivera
tiveram
tivéramos
tiverem
tivermos
tivesse
tivessem
tivéssemos
todos
trabalhar
trabalho
tu
tua
tuas
último
um
uma
umas
uns
usa
usar
valor
veja
ver
verdade
verdadeiro
você
vocês
vos
tivera
tiveram
tivéramos
tiverem
tivermos
tivesse
tivessem
tivéssemos
todos
trabalhar
trabalho
tu
tua
tuas
último
um
uma
umas
uns
usa
usar
valor
veja
ver

A.3 Lista de *Stopwords* Específica para Contexto Parlamentar

a
ahhhh
ahhh
ahh
ah
aplausos
artigo
asneiras
abstenções
baixeza
bancada
BE
bem
blá
CDS-PP
qual
Queira
certeza
coitadinho
Com
concluir
conclua
contra
contestação
contesta
contesto
de
deputado
deputada
Demagogia
do
e
É
esclarecimentos
embaraçado
Estou
Exactamente
está
Exacto
faça
fazer
favor
Finalmente
facto
fala
Falo
falou
Falso
Foram
Governo
houvesse
há
intervenção
isso
Leia
lhe
lembrado
lei
Não
nada
mais
más
Mau
mentira
Miseráveis
Ministro
Muito
o
Oh
Ora
ora
ouça
ouvi
palavra
Para
Pausa
Peço
pedir
PCP
pediu
Pensei
Pergunto
Presidente
Primeiro
Protestos
proposta
responder
Portugal
português
portugueses
PS
PSD
PCP
PP
que
Querem
Tal
Tem
terminar
responder
ridículo
risos
Roda
roda
Secretária
Secretário
Sr
Sra
Sr.ª
Sr.as
vamos
verdade
Verdes
vergonha
votam
votar
votação
voto
Vozes
vou
um
Uma
Zero

A.4 Lista de *Stopwords* Geral

acerca
acordo
afirma
afirmou
agora
ainda
alem
algumas
alguns
altura
ano
anos
antes
antonio
ali
ambos
antes
ao
aos
apenas
apesar
aplauso
aplausos
apoio
apos
apontar
aquela
aquelas
aquele
aqueles
aqui
aquilo
area
as
as
assim
associacao
assunto
ate
atras
aumento
atraves
banco
bem
bilhoes
bom
cada
caminho
camara
campanha
candidato
capital
carlos
num
casa
caso
causa
cento
central
centro
cerca
cidade
cinco
cinema
cima
coisa
com
comissao
como
congresso
conselho
conta
contos
contra
cultura
comprido
conhecido
corrente
da
dar
das
data
de
decisao
depois
deputado
debaixo
dela
delas
dele
deles
dentro
depois
desde
desligado
deve
devem
devera
desta
deste
dia
dias
dinheiro
direito
director
direccao
disse
pontos
direita
diz
dizer
do
dois
dos
duas
durante
economica
e
e
ela
elas
ele
eles
economia
eleicoes
em
embora
empresa
empresas
enquanto
entanto
entao
entre
equipa
era
eram
eramos
essa
essas
esse
esses
esta
esta
estado
estados
estamos
estao
estar
estara
estas
estava
estavam
estavamos
este
estes
eu
esteja
estejam
estejamos
estes
direita
sao
estive
estivemos
estiver
estivera
estiveram
estiveramos
estiverem
estivermos
estivesse
estivessem
estivessemos
especial
europa
europeia
exemplo
estou
eu
facto
falta
fara
faz
fazer
fernando
fazia
fez
ficou
filho
filme
fim
final
foi
folha
forma
fomos
foram
foramos
forem
formos
fosse
fossem
fossemos
frente
fui
geral
governo
grande
grandes
grupo
guerra
ha
havia
haja
hajam
srs
hajamos
hao
havemos
hei
horas
houve
houvemos
houver
houvera
houvera
houveram
houveramos
houverao
houverei
houverem
houveremos
houveria
houveriam
houveriamos
houvermos
houvesse
houvessem
houvessemos
historia
hoje
homem
iniciar
inflaccao
inicio
internacional
ir
ira
isso
ista
iste
isto
ja
janeiro
joao
jogo
jorge
jose
juros
justica
la
lado
lei
ligado
livro
lhe
lisboa
local
lugar
toda
lhes
maior
maria
maioria
maiorias
mais
manuel
mas
me
meio
melhor
menos
mercado
mes
meses
mesma
mesmo
mil
milhoes
minha
ministerio
ministro
momento
momentos
muito
muitos
mulher
mundo
musica
meu
meus
minha
minhas
muito
muitos
na
nacional
nada
nao
nas
nem
neste
no
noite
nome
nos
nos
nova
nossa
nossas
nosso
nossos
novo

numa
numero
nunca
num
numa
o
obras
onde
ontem
os
ou
outra
outras
outro
outros
p.
pais
paises
palavra
para
parece
parte
pegar
pela
pelas
pelo
pelos
pessoas
pode
podera
podia
por
porque
povo
paulo
promeiro
partido
partir
passado
pela
pelas
pelo
pelos
pesquisa
pesquisas
pessoas
plano
pode
poder
podera
policia
politica
por
porque
porto
portanto
portugal
portugues
portuguesa
portugueses
possivel
pouco
presidente
primeira
primeiro
problema
problemas
processo
producao
produtos
programa
projecto
proprio
pt
proximo
ps
psd
publico
quais
qual
qualquer
quando
quanto
quase
quatro
que
quem
quer
questao
qual
que
que
quem
quieto
r
r.
real
recursos
regiao
relacao
republica
reportagem
rio
saber
saude
se
seja
sejam
sejamos
sem
semana
sempre
sendo
ser
sera
serao
serei
seremos
seria
serie
seriea
serieb
seriec
seried
seriam
seriamos
seu
seus
segunda
segundo
seguranca
seis
seja
sem
semana
sempre
sentido
ser
sera
seu
seus
sido
silva
sistema
situacao
so
sobre
social
sociedade
sua
suas
Sr
Srs
Sra
Sras
sr
sra
sras
somente
somos
sou
sua
suas
sucursal
sul
tal
tambem
tao
tarde
te
tem
tem
tem
temos
tempo
tenha
tenham
tenhamos
tenho
tentar
tentaram
tente
tentei
tera
terao
terei
teremos
teria
teriam
teriamos
teu
teus
teve
tinha
tinham
tinhamos
tipo
tive
tivemos
tiver
tivera
tiveram
tiveramos
tiverem
tivermos
tivesse
tivessem
tivessemos
todas
todo
todos
trabalhar
trabalho
tres
tu
tudo
tua
tuas
ultimo
ultimos
um
uma
umas
uns
usa
usar
vai
vao
ver
vez
vezes
vida
valor
veja
ver
verdade
verdadeiro
voce
voces
vos
zona
vos
toda
todas
todo
todos
trabalhar
trabalho
tres
tu
tudo
tua
tuas
ultimo
ultimos
um
uma
umas
uns
usa

Glossário

AR - Assembleia da República

CDS-PP - Partido Popular

DAR – Diário da Assembleia da Republica

Eurovoc **-** Lista de termos (*Thesaurus*) multilingue que cobre todos os domínios da actividade das comunidades

KDD (*Knowlegde Discovery in Databases*) - descoberta do conhecimento em bases de dados

Knn (*k-nearest neighbors algorithm*) é um algoritmo de classificação que pode ser aplicado a varios formatos de dados.

KnnFlex – Algoritmos de classificação para textos

Mining – mineração

NLP (*Natural Language processing*) é uma área da ciência do computador e linguística que investiga a interacção entre computador e linguagem humana.

OLAP (*On-line Analytical Processing)* - é a capacidade para manipular e analisar um grande volume de dados sob múltiplas perspectivas.

Open source – Software de código aberto

openNLP – Aplicação de mineração de texto para processamento da linguagem natural

PCP - Partido Comunista Português

PDF (*Portable Document Format*) - é um formato de ficheiro, desenvolvido pela Adobe Systems, para representar documentos criados independentemente da aplicação, do *hardware* e do sistema operativo.

PLN- Processamento da linguagem Natural

PS – Partido socialista

PSD - Partido Social Democrata

Stopwords – lista de palavras que não acrescentam nenhum significado adicional ao conteúdo dum texto

SVM (*support vector machines*) – algoritmo de classificação que pode ser aplicado a vários formatos de dados

Tm - *text mining*

TXT – formato dum ficheiro que contém texto pleno, existente dentro do sistema de ficheiro do computador.

Web – documentos retirados na internet ou *World Wide Web* que é um sistema de documentos em hipermédia que são interligados e executados na Internet.

XML - (*eXtensible Markup Language*) é formato dum tipo de ficheiro que surgiu a partir duma recomendação da W3C para gerar linguagens de marcação para necessidades especiais. Esta linguagem é capaz de descrever diversos tipos de dados e facilitar a partilha de informações através da Internet.